Rajiv Khosla
Robert J. Howlett
Lakhmi C. Jain (Eds.)

LNAI 3683

Knowledge-Based Intelligent Information and Engineering Systems

9th International Conference, KES 2005
Melbourne, Australia, September 2005
Proceedings, Part III

3 Part III

Springer

Lecture Notes in Artificial Intelligence 3683

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Rajiv Khosla Robert J. Howlett
Lakhmi C. Jain (Eds.)

Knowledge-Based Intelligent Information and Engineering Systems

9th International Conference, KES 2005
Melbourne, Australia, September 14-16, 2005
Proceedings, Part III

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Rajiv Khosla
La Trobe University
Business Systems and Knowledge Modelling Laboratory, School of Business
Victoria 3086, Australia
E-mail: R.Khosla@latrobe.edu.au

Robert J. Howlett
University of Brighton
School of Engineering, Engineering Research Centre
Moulsecoomb, Brighton, BN2 4GJ, UK
E-mail: r.j.howlett@bton.ac.uk

Lakhmi C. Jain
University of South Australia
School of Electrical and Information Engineering, KES Centre
Mawson Lakes Campus, Adelaide, South Australia SA 5095, Australia
E-mail: Lakhmi.Jain@unisa.edu.au

Library of Congress Control Number: 2005932202

CR Subject Classification (1998): I.2, H.4, H.3, J.1, H.5, K.6, K.4

ISSN 0302-9743
ISBN-10 3-540-28896-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-28896-1 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 11553939 06/3142 5 4 3 2 1 0

Preface

Dear delegates, friends and members of the growing KES professional community, welcome to the proceedings of the 9th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems hosted by La Trobe University in Melbourne Australia.

The KES conference series has been established for almost a decade, and it continues each year to attract participants from all geographical areas of the world, including Europe, the Americas, Australasia and the Pacific Rim. The KES conferences cover a wide range of intelligent systems topics. The broad focus of the conference series is the theory and applications of intelligent systems. From a pure research field, intelligent systems have advanced to the point where their abilities have been incorporated into many business and engineering application areas. KES 2005 provided a valuable mechanism for delegates to obtain an extensive view of the latest research into a range of intelligent-systems algorithms, tools and techniques. The conference also gave delegates the chance to come into contact with those applying intelligent systems in diverse commercial areas. The combination of theory and practice represented a unique opportunity to gain an appreciation of the full spectrum of leading-edge intelligent-systems activity.

The papers for KES 2005 were either submitted to invited sessions, chaired and organized by respected experts in their fields, or to a general session, managed by an extensive International Program Committee, or to the Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP) Workshop, managed by an International Workshop Technical Committee. Whichever route they came through, all papers for KES 2005 were thoroughly reviewed. The adoption by KES of the PROSE Publication Review and Organisation System software greatly helped to improve the transparency of the review process and aided quality control.

In total, 1382 papers were submitted for KES 2005, and a total of 688 papers were accepted, giving an acceptance rate of just under 50%. The proceedings, published this year by Springer, run to more than 5000 pages. The invited sessions are a valuable feature of KES conferences, enabling leading researchers to initiate sessions that focus on innovative new areas. A number of sessions in new emerging areas were introduced this year, including Experience Management, Emotional Intelligence, and Smart Systems. The diversity of the papers can be judged from the fact that there were about 100 technical sessions in the conference program. More than 400 universities worldwide participated in the conference making it one of the largest conferences in the area of intelligent systems. As would be expected, there was good local support with the participation of 20 Australian universities. There was a significant business presence, provided by the involvement of a number of industry bodies, for example, CSIRO Australia, DSTO Australia, Daewoo South Korea and NTT Japan.

KES International gratefully acknowledges the support provided by La Trobe University in hosting this conference. We acknowledge the active interest and support from La Trobe University's Vice Chancellor and President, Prof. Michael Osborne, Dean of the Faculty of Law and Management, Prof. Raymond Harbridge, Dean of the Faculty of Science and Technology, Prof. David Finlay, and Head of the School of Business, Prof. Malcolm Rimmer. KES International also gratefully acknowledges the support provided by Emeritus Prof. Greg O'Brien.

A tremendous amount of time and effort goes into the organization of a conference of the size of KES 2005. The KES community owes a considerable debt of gratitude to the General Chair Prof. Rajiv Khosla and the organizing team at La Trobe University for their huge efforts this year in bringing the conference to a successful conclusion. As the conference increases in size each year the organizational effort needed increases and we would like to thank Prof. Khosla and his colleagues for coping efficiently with the largest KES conference to date.

We would like to thank the Invited Session Chairs, under the leadership and guidance of Prof. Lakhmi Jain and Prof. Rajiv Khosla for producing high-quality sessions on leading-edge topics. We would like to thank the KES 2005 International Program Committee for undertaking the considerable task of reviewing all of the papers submitted for the conference. We express our gratitude to the high-profile keynote speakers for providing talks on leading-edge topics to inform and enthuse our delegates. A conference cannot run without authors to write papers. We thank the authors, presenters and delegates to KES 2005 without whom the conference could not have taken place. Finally we thank the administrators, caterers, hoteliers, and the people of Melbourne for welcoming us and providing for the conference.

We hope you found KES 2005 a worthwhile, informative and enjoyable experience.

July 2005

Bob Howlett
Rajiv Khosla
Lakhmi Jain

KES 2005 Conference Organization

General Chair

Rajiv Khosla
Business Systems and Knowledge Modelling Laboratory
School of Business
La Trobe University
Melbourne, Victoria 3086
Australia

Conference Founder and Honorary Program Committee Chair

Lakhmi C. Jain
Knowledge-Based Intelligent Information and Engineering Systems Centre
University of South Australia, Australia

KES Executive Chair

Bob Howlett
Intelligent Systems and Signal Processing Laboratories/KTP Centre
University of Brighton, UK

KES Journal General Editor

Bogdan Gabrys
University of Bournemouth, UK

Local Organizing Committee

Malcolm Rimmer – Chair, School of Business, La Trobe University
Rajiv Khosla, Selena Lim, Brigitte Carrucan, Monica Hodgkinson, Marie Fenton, Maggie Van Tonder, and Stephen Muir
La Trobe University, Melbourne, Australia

KES 2005 Web Page Design Team

Joe Hayes, Anil Varkey Samuel, Mehul Bhatt, Rajiv Khosla
La Trobe University, Melbourne, Australia

KES 2005 Liaison and Registration Team

Rajiv Khosla, Selena Lim, Brigitte Carrucan, Jodie Kennedy, Maggie Van Tonder, Marie Fenton, Colleen Stoate, Diane Kraal, Cary Slater, Petrus Usmanij, Chris Lai, Rani Thanacoody, Elisabeth Tanusasmita, George Plocinski
La Trobe University, Melbourne, Australia

KES 2005 Proceedings Assembly Team

Rajiv Khosla
Selena Lim
Monica Hodgkinson
George Plocinski
Maggie Van Tonder
Colleen Stoate
Anil Varkey Samuel
Marie Fenton
Mehul Bhatt
Chris Lai
Petrus Usmanij

International Program Committee

Hussein Abbass, University of New South Wales, Australia
Akira Asano, Hiroshima University, Japan
Robert Babuska, Delft University of Technology, The Netherlands
Patrick Bosc, IRISA/ENSSAT, France
Pascal Bouvry, Luxembourg University of Applied Sciences, Luxembourg
Krzysztof Cios, University of Colorado, USA
Carlos A. Coello, LANIA, Mexico
Ernesto Damiani, University of Milan, Italy
Da Deng, University of Otago, New Zealand
Vladan Devedzic, University of Belgrade, Serbia and Montenegro
Vladimir Gorodetski, Russian Academy of Sciences, Russia
Manuel Grana, UPV/EHU, Spain
Lars Kai Hansen, Technical University of Denmark, Denmark
Yuzo Hirai, University of Tsukuba, Japan
Daniel Howard, QinetiQ, UK
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Hisao Ishibuchi, Osaka Prefecture University, Japan
Naohiro Ishii, Aichi Institute of Technology, Japan
Seong-Joon Yoo, Sejong University, Korea
Nikos Karacapilidis, University of Patras, Greece
Rajiv Khosla, La Trobe University, Australia
Laszlo T. Koczy, Budapest and Szechenyi Istvan University, Hungary
Andrew Kusiak, University of Iowa, USA
W.K. Lai, MIMOS Berhad, Malaysia
Ian C. Parmee, University of the West of England, UK
Dong Hwa Kim, Hanbat National University, Korea
Jingli Cao, La Trobe University, Australia
Da Ruan, Belgian Nuclear Research Centre, Belgium
Ang Yang, University of New South Wales, Australia
Adrian Stoica, NASA Jet Propulsion Laboratory, USA
Janusz Kacprzyk, Polish Academy of Sciences, Poland
Vijayan Asari, Old Dominion University, USA
Raymond Lee, Hong Kong Polytechnic University, Hong Kong, China
Chee-Peng Lim, University of Science, Malaysia
Ignac Lovrek, University of Zagreb, Croatia
Bob McKay, University of NSW, Australia
Dan C. Marinescu, University of Central Florida, USA
Radko Mesiar, Slovak Technical University, Slovakia
Hirofumi Nagashino, University of Tokushima, Japan
Ciprian Daniel Neagu, University of Bradford, UK
Ngoc Thanh Nguyen, Wroclaw University of Technology, Poland

Toyoaki Nishida, University of Tokyo, Japan
Vasile Palade, Oxford University, UK
John A. Rose, University of Tokyo, Japan
Rajkumar Roy, Cranfield University, UK
Udo Seiffert, Leibniz Institute of Plant Genetics and Crop Plant Research, Germany
Flavio Soares Correa da Silva, University of Sao Paulo, Brazil
Von-Wun Soo, National Tsing Hua University, Taiwan
Sarawut Sujitjorn, Suranaree University of Technology, Thailand
Takushi Tanaka, Fukuoka Institute of Technology, Japan
Eiichiro Tazaki, University of Yokohama, Japan
Jon Timmis, University of Kent, UK
Jim Torresen, University of Oslo, Norway
Andy M. Tyrrell, University of York, UK
Eiji Uchino, University of Yamaguchi, Japan
Jose Luis Verdegay, University of Granada, Spain
Dianhui Wang, La Trobe University, Australia
Junzo Watada, Waseda University, Japan
Xin Yao, University of Birmingham, UK
Boris Galitsky, University of London, UK
Bogdan Gabrys, Bournemouth University, UK
Norbert Jesse, Universität Dortmund, Germany
Keiichi Horio, Kyushu Institute of Technology, Japan
Bernd Reusch, University of Dortmund, Germany
Nadia Berthouze, University of Aizu, Japan
Hideyaki Sawada, Kagawa University, Japan
Yasue Mitsukura, University of Okayama, Japan
Dharmendra Sharma, University of Canberra, Australia
Adam Grzech, Wroclaw University of Technology, Poland
Mircea Negoita, Wellington Institute of Technology, New Zealand
Hongen Lu, La Trobe University, Australia
Kosuke Sekiyama, University of Fukui, Japan
Raquel Flórez-López, Campus de Vegazana, Spain
Eugénio Oliveira, University of Porto, Portugal
Roberto Frias, University of Porto, Portugal
Maria Virvou, University of Piraeus, Greece
Daniela Zambarbieri, Università di Pavia, Italy
Andrew Skabar, La Trobe University, Australia
Zhaohao Sun, University of Wollongong, Australia
Koren Ward, University of Wollongong, Australia

Invited Session Chairs Committee

Krzysztof Cios, University of Denver, USA
Toyoaki Nishida, University of Tokyo, Japan
Ngoc Thanh Nguyen, Wroclaw University of Technology, Poland
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Yoshinori Adachi, Chubu University, Japan
Nobuhiro Inuzuka, Nagoya Institute of Technology, Japan
Naohiro Ishii, Aichii Institute of Technology, Japan
Yuji Iwahori, Chubu University, Japan
Tai-hoon Kim, Security Engineering Research Group (SERG), Korea
Daniel Howard, QinetiQ, UK
Mircea Gh. Negoita, Wellington Institute of Technology, New Zealand
Akira Suyama, University of Tokyo, Japan
Dong Hwa Kim, Hanbat National University, Korea
Hussein A. Abbas, University of New South Wales, Australia
William Grosky, University of Michigan-Dearborn, USA
Elisa Bertino, Purdue University, USA
Maria L. Damiani, University of Milan, Italy
Kosuke Sekiyama, University of Fukui, Japan
Seong-Joon Yoo, Sejong Univwersity, Korea
Hirokazu Taki, Wakayama University, Japan
Satoshi Hori, Institute of Technologists, Japan
Hideyuki Sawada, Kagawa University, Japan
Vicenc Torra, Artificial Intelligence Research Institute, Spain
Maria Vamrell, Artificial Intelligence Research Institute, Spain
Manfred Schmitt, Technical University of Munich, Germany
Denis Helic, Graz University of Technology, Austria
Yoshiteru Ishida, Toyohashi University of Technology, Japan
Giuseppina Passiante, University of Lecce, Italy
Ernesto Damiani, University of Milan, Italy
Susumu Kunifuji, Japan Advanced Institute of Science and Technology, Japan
Motoki Miura, Japan Advanced Institute of Science and Technology, Japan
James Liu, Hong Kong Polytechnic University, Hong Kong, China
Honghua Dai, Deakin University, Australia
Pascal Bouvry, Luxembourg University, Luxembourg
Gregoire Danoy, Luxembourg University, Luxembourg
Ojoung Kwon, California State University, USA
Jun Munemori, Wakayama University, Japan
Takashi Yoshino, Wakayama University, Japan
Takaya Yuizono, Shimane University, Japan
Behrouz Homayoun Far, University of Calgary, Canada
Toyohide Watanabe, Nagoya University, Japan
Dharmendra Sharma, University of Canberra, Australia
Dat Tran, University of Canberra, Australia

Kim Le, University of Canberra, Australia
Takumi Ichimura, Hiroshima City University, Japan
K Yoshida, St. Marianna University, Japan
Phill Kyu Rhee, Inha University, Korea
Chong Ho Lee, Inha University, Korea
Mikhail Prokopenko, CSIRO ICT Centre, Australia
Daniel Polani, University of Hertfordshire, UK
Dong Chun Lee, Howon University, Korea
Dawn E. Holmes, University of California, USA
Kok-Leong Ong, Deakin University, Australia
Vincent Lee, Monash University, Australia
Wee-Keong Ng, Nanyang Technological University
Gwi-Tae Park, Korea University, Korea
Giles Oatley, University of Sunderland, UK
Sangkyun Kim, Korea Information Engineering Service, Korea
Hong Joo Lee, Daewoo Electronics Corporation, Korea
Ryohei Nakano, Nagoya Institute of Technology, Japan
Kazumi Saito, NTT Communication Science Laboratories, Japan
Kazuhiko Tsuda, University of Tsukuba, Japan
Torbjørn Nordgård, Norwegian University of Science and Technology, Norway
Øystein Nytrø, Norwegian University of Science and Technology, Norway
Amund Tveit, Norwegian University of Science and Technology, Norway
Thomas Brox Røst, Norwegian University of Science and Technology, Norway
Manuel Graña, Universidad Pais Vasco, Spain
Richard Duro, Universidad de A Coruña, Spain
Kazumi Nakamatsu, University of Hyogo, Japan
Jair Minoro Abe, University of Sao Paulo, Brazil
Hiroko Shoji, Chuo University, Japan
Yukio Ohsawa, University of Tsukuba, Japan
Da Deng, University of Otago, New Zealand
Irena Koprinska, University of Sydney, Australia
Eiichiro Tazaki, University of Yokohama, Japan
Kenneth J. Mackin, Tokyo University of Information Sciences, Japan
Lakhmi Jain, University of South Australia, Australia
Tetsuo Fuchino, Tokyo Institute of Technology, Japan
Yoshiyuki Yamashita, Tohoku University, Japan
Martin Purvis, University of Otago, New Zealand
Mariusz Nowostawski, University of Otago, New Zealand
Bastin Tony Roy Savarimuthu, University of Otago, New Zealand
Norio Baba, Osaka Kyoiku University, Japan
Junzo Watada, Waseda University, Japan
Petra Povalej, Laboratory for System Design, Slovenia
Peter Kokol, Laboratory for System Design, Slovenia
Woochun Jun, Seoul National University of Education, Korea
Andrew Kusiak, University of Iowa, USA
Hanh Pham, State University of New York, USA

IIHMSP Workshop Organization Committee

General Co-chairs

Jeng-Shyang Pan
National Kaohsiung University of Applied Sciences, Taiwan
Lakhmi C. Jain
University of South Australia, Australia

Program Committee Co-chairs

Wai-Chi Fang
California Institute of Technology, USA
Eiji Kawaguchi
Keio University, Japan

Finance Chair

Jui-Fang Chang
National Kaohsiung University of Applied Sciences, Taiwan

Publicity Chair

Kang K. Yen
Florida International University, USA

Registration Chair

Yan Shi
Tokai University, Japan

Electronic Media Chair

Bin-Yih Liao
National Kaohsiung University of Applied Sciences, Taiwan

Publication Chair

Hsiang-Cheh Huang
National Chiao Tung University, Taiwan

Local Organizing Chair

R. Khosla
La Trobe University, Australia

Asia Liaison

Yoiti Suzuki
Tohoku University, Japan

North America Liaison

Yun Q. Shi
New Jersey Institute of Technology, USA

Europe Liaison

R.J. Howlett
University of Brighton, UK

IIHMSP Workshop Technical Committee

Prof. Oscar Au, Hong Kong University of Science and Technology, Hong Kong, China
Prof. Chin-Chen Chang, National Chung Cheng University, Taiwan
Prof. Chang Wen Chen, Florida Institute of Technology, USA
Prof. Guanrong Chen, City University of Hong Kong, Hong Kong, China
Prof. Liang-Gee Chen, National Taiwan University, Taiwan
Prof. Tsuhan Chen, Carnegie Mellon University, USA
Prof. Yung-Chang Chen, National Tsing-Hua University, Taiwan
Prof. Yen-Wei Chen, Ritsumeikan University, Japan
Prof. Hsin-Chia Fu, National Chiao Tung University, Taiwan
Prof. Hsueh-Ming Hang, National Chiao Tung University, Taiwan
Prof. Dimitrios Hatzinakos, University of Toronto, Canada
Dr. Ichiro Kuroda, NEC Electronics Corporation, Japan

KES 2005 Reviewers

H. Abbass, University of New South Wales, Australia
J.M. Abe, University of Sao Paulo, Brazil
Y. Adachi, Chubu University, Japan
F. Alpaslan, Middle East Technical University, Turkey
P. Andreae, Victoria University, New Zealand
A. Asano, Hiroshima University, Japan
K.V. Asari, Old Dominion University, USA
N. Baba, Osaka-Kyoiku University, Japan
R. Babuska, Delft University of Technology, The Netherlands
P. Bajaj, G.H. Raisoni College of Engineering, India
A. Bargiela, Nottingham Trent University, UK
M. Bazu, Institute of Microtechnology, Romania
N. Berthouze, University of Aizu, Japan
E. Bertino, Purdue University, USA
Y. Bodyanskiy, Kharkiv National University of Radioelectronics, Ukraine
P. Bosc, IRISA/ENSSAT, France
P. Bouvry, Luxembourg University, Luxembourg
P. Burrell, South Bank University, UK
J. Cao, La Trobe University, Australia
B. Chakraborty, Iwate Prefectural University, Japan
Y.-W. Chen, Ryukyus University, Japan
Y.-H. Chen-Burger, University of Edinburgh, UK
V. Cherkassky, University of Minnesota, USA
K. Cios, University at Denver, USA
C.A. Coello, LANIA, Mexico
G. Coghill, University of Auckland, New Zealand
D. Corbett, SAIC, USA
D.W. Corne, University of Exeter, UK
D. Cornforth, Charles Sturt University, Australia
F.S.C. da Silva, University of Sao Paulo, Brazil
H. Dai, Deakin University, Australia
E. Damiani, University of Milan, Italy
M.L. Damiani, University of Milan, Italy
G. Danoy, Luxembourg University, Luxembourg
K. Deep, Indian Institute of Technology Roorkee, India
D. Deng, University of Otago, New Zealand
V. Devedzic, University of Belgrade, Serbia and Montenegro
D. Dubois, Université Paul Sabatier, France
R. Duro, Universidad de A Coruña, Spain
D. Earl, Oak Ridge National Laboratory, USA
B. Far, University of Calgary, Canada

M. Fathi, National Magnet Laboratory, USA
R. Flórez-López, Campus de Vegazana, Spain
M. Frean, Victoria University of Wellington, New Zealand
R. Frias, University of Porto, Portugal
T. Fuchino, Tokyo Institute of Technology, Japan
P. Funk, Mälardalen University, Sweden
B. Gabrys, University of Bournemouth, UK
B. Galitsky, University of London, UK
T. Gedeon, Murdoch University, Australia
M. Gen, Waseda University, Japan
A. Ghosh, ISICAI, India
V. Gorodetski, Russian Academy of Sciences, Russia
M. Grana, Facultad de Informatica UPV/EHU, Spain
W. Grosky, University of Michigan-Dearborn, USA
A. Grzech, Wroclaw University of Technology, Poland
D. Gwaltney, NASA George C. Marshall Space Flight Center, USA
L.K. Hansen, Technical University of Denmark, Denmark
C.J. Harris, University of Southampton, UK
D. Helic, Graz University of Technology Austria
L. Hildebrand, University of Dortmund, Germany
Y. Hirai, University of Tsukuba, Japan
D.E. Holmes, University of California, USA
B. Homayoun, Far University of Calgary, Canada
T.-P. Hong, National University of Kaohsiung, Taiwan
S. Hori, Institute of Technologists, Japan
K. Horio, Kyushu Institute of Technology, Japan
D. Howard, QinetiQ, UK
B. Howlett, University of Brighton, UK
M.-P. Huget, University of Savoie, France
H. Iba, University of Tokyo, Japan
T. Ichimura, Hiroshima City University, Japan
N. Inuzuka, Nagoya Institute of Technology, Japan
H. Ishibuchi, Osaka Prefecture University, Japan
Y. Ishida, Toyohashi University of Technology, Japan
N. Ishii, Aichi Institute of Technology, Japan
Y. Iwahori, Chubu University, Japan
L. Jain, University of South Australia, Australia
M.M. Jamshidi, University of New Mexico, USA
N. Jesse, Universität Dortmund, Germany
S. Joo, Sejong University, Korea
W. Jun, Seoul National University of Education, Korea
J. Kacprzyk, Polish Academy of Sciences, Poland
N. Karacapilidis, University of Patras, Greece
V. Kecman, Auckland University, New Zealand
R. Khosla, La Trobe University, Australia

D.H. Kim, Hanbat National University, Korea
S. Kim, Korea Information Engineering Service, Korea
T.-H. Kim, Security Engineering Research Group (SERG), Korea
L.T. Koczy, Budapest University of Technology and Economics, Hungary
P. Kokol, Laboratory for System Design, Slovenia
A. Konar, Jadavpur University, India
I. Koprinska, University of Sydney, Australia
H. Koshimizu, Chukyo University, Japan
S. Kunifuji, Japan Advanced Institute of Science and and Technology, Japan
A. Kusiak, University of Iowa, USA
O. Kwon, California State University, USA
W.K. Lai, MIMOS Berhad, Malaysia
P.L. Lanzi, Polytechnic Institute, Italy
K. Le, University of Canberra, Australia
C.H. Lee, Inha University, Korea
D.C. Lee, Howon University, Korea
H.J. Lee, Daewoo Electronics Corporation, Korea
R. Lee, Hong Kong Polytechnic University, Hong Kong, China
V. Lee, Monash University, Australia
Q. Li, La Trobe University, Australia
C.-P. Lim, University of Science, Malaysia
S. Lim, La Trobe University, Australia
J. Liu, Hong Kong Polytechnic University, Hong Kong, China
I. Lovrek, University of Zagreb, Croatia
H. Lu, La Trobe University, Australia
B. MacDonald, Auckland University, New Zealand
K.J. Mackin, Tokyo University of Information Sciences, Japan
L. Magdalena-Layos, EUSFLAT, Spain
D.C. Marinescu, University of Central Florida, USA
F. Masulli, University of Pisa, Italy
J. Mazumdar, University of South Australia, Australia
B. McKay, University of NSW, Australia
S. McKinlay, Wellington Institute of Technology, New Zealand
R. Mesiar, Slovak Technical University, Slovakia
J. Mira, UNED, Spain
Y. Mitsukura, University of Okayama, Japan
M. Miura, Japan Advanced Institute of Science and Technology, Japan
J. Munemori, Wakayama University, Japan
H. Nagashino, University of Tokushima, Japan
N. Nagata, Chukyo University, Japan
K. Nakajima, Tohoku University, Japan
K. Nakamatsu, University of Hyogo, Japan
R. Nakano, Nagoya Institute, Japan
T. Nakashima, Osaka University, Japan
L. Narasimhan, University of Newcastle, Australia

V.E. Neagoe, Technical University, Romania
C.D. Neagu, University of Bradford, UK
M.G. Negoita, WelTec, New Zealand
W.-K. Ng, Nanyang Technological University, Singapore
C. Nguyen, Catholic University of America, USA
N.T. Nguyen, Wroclaw University of Technology, Poland
T. Nishida, University of Tokyo, Japan
T. Nordgård, Norwegian University of Science and Technology, Norway
M. Nowostawski, University of Otago, New Zealand
Ø. Nytrø, Norwegian University of Science and Technology, Norway
G. Oatley, University of Sunderland, UK
Y. Ohsawa, University of Tsukuba, Japan
E. Oliveira, University of Porto, Portugal
K.-L. Ong, Deakin University, Australia
N.R. Pal, Indian Statistical Institute, India
V. Palade, Oxford University, UK
G.-T. Park, Korea University, Korea
I.C. Parmee, University of the West of England, UK
G. Passiante, University of Lecce, Italy
C.-A. Peña-Reyes, Swiss Federal Institute of Technology - EPFL, Switzerland
H. Pham, State University of New York, USA
D. Polani, University of Hertfordshire, UK
T. Popescu, National Institute for Research and Development Informatic, Italy
P. Povalej, Laboratory for System Design, Slovenia
M. Prokopenko, CSIRO ICT Centre, Australia
M. Purvis, University of Otago, New Zealand
G. Resconi, Catholic University, Italy
B. Reusch, University of Dortmund, Germany
P.K. Rhee, Inha University, Korea
J.A. Rose, University of Tokyo, Japan
T.B. Røst, Norwegian University of Science and Technology, Norway
E. Roventa, York University, Canada
R. Roy, Cranfield University, UK
D. Ruan, Belgian Nuclear Research Centre, Belgium
A. Saha, NCD, Papua New Guinea
K. Saito, NTT Communication Science Laboratories, Japan
T. Samatsu, Kyushu Tokai University, Japan
E. Sanchez, Université de la Méditeranée, France
B.T.R. Savarimuthu, University of Otago, New Zealand
H. Sawada, Kagawa University, Japan
M. Schmitt, Technical University of Munich, Germany
M. Schoenauer, INRIA, France
U. Seiffert, Leibniz Institute of Plant Genetics
and Crop Plant Research Gatersleben, Germany
K. Sekiyama, University of Fukui, Japan

D. Sharma, University of Canberra, Australia
H. Shoji, Chuo University, Japan
A. Skabar, La Trobe University, Australia
B. Smyth, University College Dublin, Ireland
V.-W. Soo, National Tsing Hua University, Taiwan
A. Stoica, NASA Propulsion Jet Laboratory, USA
M.R. Stytz, Yamaguchi University, Japan
N. Suetake, Yamaguchi University, Japan
S. Sujitjorn, Suranaree University of Technology, Thailand
Z. Sun, University of Wollongong, Australia
A. Suyama, University of Tokyo, Japan
H. Taki, Wakayama University, Japan
T. Tanaka, Fukuoka Institute of Technology, Japan
M. Tanaka-Yamawaki, Tottori University, Japan
E. Tazaki, University of Yokohama, Japan
S. Thatcher, University of South Australia, Australia
P. Theodor, National Institute for Research and Development Informatics, Romania
J. Timmis, University of Kent at Canterbury, UK
V. Torra, Artificial Intelligence Research Institute, Spain
J. Torresen, University of Oslo, Norway
D. Tran, University of Canberra, Australia
K. Tsuda, University of Tsukuba, Japan
C. Turchetti, Università Politecnica delle Marche, Italy
A. Tveit, Norwegian University of Science and Technology, Norway
J. Tweedale, Defence Science and Technology Organization, Australia
A.M. Tyrrell, University of York, UK
E. Uchino, University of Yamaguchi, Japan
A. Uncini, University of Rome, Italy
P. Urlings, Defence Science and Technology Organization, Australia
M. Vamrell, Artificial Intelligence Research Institute, Spain
J.L. Verdegay, University of Granada, Spain
M. Virvou, University of Piraeus, Greece
S. Walters, University of Brighton, UK
D. Wang, La Trobe University, Australia
L. Wang, Nanyang Technical University, Singapore
P. Wang, Temple University, USA
K. Ward, University of Wollongong, Australia
J. Watada, Waseda University, Japan
K. Watanabe, Saga University, Japan
T. Watanabe, Nagoya University, Japan
T. Yamakawa, Kyushu Institute of Technology, Japan
Y. Yamashita, Tohoku University, Japan
A. Yang, University of New South Wales, Australia
X. Yao, University of Birmingham, UK
S.-J. Yoo, Sejong University, Korea

K. Yoshida, Kyushu Institute of Technology, Japan
T. Yoshino, Wakayama University, Japan
T. Yuizono, Shimane University, Japan
L. Zadeh, Berkeley University of Caliornia, USA
D. Zambarbieri, Università di Pavia, Italy
X. Zha, National Institute of Standards and Technology, USA

KES 2005 Keynote Speakers

1. **Professor Jun Liu**, Harvard University, MA, USA
   Topic: From Sequence Information to Gene Expression

2. **Professor Ron Sun**, Rensselaer Polytechnic Institute, New York, USA
   Topic: From Hybrid Systems to Hybrid Cognitive Architectures

3. **Professor Jiming Liu**, Hong Kong Baptist University, Hong Kong, China
   Topic: Towards Autonomy Oriented Computing (AOC):
   Formulating Computational Systems with Autonomous Components

4. **Professor Toyoaki Nishida**, Kyoto University and Tokyo University, Japan
   Topic: Acquiring, Accumulating, Transforming, Applying,
   and Understanding Conversational Quanta

5. **Professor Marimuthu Palaniswami**, University of Melbourne, Australia
   Topic: Convergence of Smart Sensors and Sensor Networks

Table of Contents, Part III

Intelligent Agent Ontologies and Environments

Intelligent Multimedia Solutions and the Security in the Next Generation Mobile Networks

Intelligent E-Mail Analysis, News Extraction and Web Mining

Semantic Integration and Ontologies

Computer Vision, Image Processing and Retrieval

Communicative Intelligence II

Approaches and Methods to Security Engineering V

Multimedia Retrieval II

Multimedia Compression

Multimedia Signal Processing

Emergence and Self-organisation in Complex Systems

Soft Computing Techniques and Their Applications III

Information Engineering and Ubiquitous Computing

Location and Context-Based Systems

e-Based Systems in Education, Commerce and Health

Computational Biology and Bioinformatics

Complex Adaptive Systems

Communicative Intelligent III

Speech Processing and Robotics

Data Mining and Soft Computing Applications I

Multimedia Security and Stegnography

Stegnography

Soft Computing Approach to Industrial Engineering II

Medical Text Mining and Natural Language Processing

Knowledge Based Intelligent Systems for Health Care

Intelligent Learning Environment

Intelligent Data Analysis and Applications

Table of Contents, Part IV

Innovations in Intelligent Systems and Their Applications

Data Mining and Soft Computing Applications II

Skill Acquisition and Ubiquitous Human Computer Interaction

Soft Computing and Their Applications – IV

Agent-Based Workflows, Knowledge Sharing and Reuse

Multi-media Authentication and Watermarking Applications

Knowledge and Engineering Techniques for Spatio-temporal Applications

Intelligent Data Analysis and Applications II

Creativitiy Support Environment and Its Social Applications

Collective Intelligence

Computational Methods for Intelligent Neuro-fuzzy Applications

Evolutionary and Self-organizing Sensors, Actuators and Processing Hardware

Knowledge Based Systems for e-Business and e-Learning I

Multi-agent Systems and Evolutionary Computing

Ubiquitous Pattern Recognition

Neural Networks for Data Mining

Intelligent Systems for e-Business and e-Learning II

Knowledge-Based Technology in Crime Matching, Modelling and Prediction

Soft Computing Applications

Table of Contents, Part I

Soft Computing Techniques in Stock Markets

Intelligent Network Based Education

Maintenance and Customization of Business Knowledge

Intelligent Data Processing in Process Systems and Plants

Intelligent Agent Technology and Applications I

Intelligent Design Support Systems

Data Engineering, Knowledge Engineering and Ontologies

Knowledge Discovery and Data Mining

Advanced Network Application

Approaches and Methods of Security Engineering I

Chance Discovery I

Information Hiding and Multimedia Signal Processing

Soft Computing Techniques and Their Applications I

Intelligent Agent Technology and Applications II

Smart Systems

Knowledge – Based Interface Systems

Intelligent Information Processing for Remote Sensing

Intelligent Human Computer Interaction Systems

Experience Management and Knowledge Management

Network (Security) Real-Time and Fault Tolerant Systems

Advanced Network Application and Real-Time Systems

Approaches and Methods of Security Engineering II

Chance Discovery II

Intelligent Watermarking Algorithms

Soft Computing Techniques and Their Applications II

Table of Contents, Part II

Machine Learning

Immunity-Based Systems

Medical Diagnosis

Intelligent Hybrid Systems and Control

Emotional Intelligence and Smart Systems

Context-Aware Evolvable Systems

Intelligant Fuzzy Systems and Control

Knowledge Representation and Its Practical Application in Today's Society

Approaches and Methods into Security Engineering III

Communicative Intelligent I

Intelligent Watermaking Algorithms and Applications

Intelligent Techniques and Control

e-Learning and ICT

Logic Based Intelligent Information Systems

Intelligent Agents and Their Applications I

Innovations in Intelligent Agents

Ontologies and the Semantic Web

Knowledge Discovery in Data Streams

Computational Intelligence Tools Techniques and Algorithms

Approaches and Methods to Security Engineering IV

Watermaking Applications I

Watermaking Applications II

Multimedia Retrieval I

Soft Computing Approach to Industrial Engineering

Experience Management and Information Systems

Agent-Based Approach for Dynamic Ontology Management

Li Li, Baolin Wu, and Yun Yang

Faculty of Information and Communication Technologies
Swinburne University of Technology
PO Box 218, Hawthorn, Melbourne, Australia 3122
{lli,bwu,yyang}@it.swin.edu.au

Abstract. In this paper, we present an ontology definition which is helpful in symbolising the ontology and ontology management. By investigating interactions between agents in business scenarios, we adopt the process algebra to model interactions and corresponding effects. For the purpose of solving the conflicts or inconsistency in understanding a particular issue, we investigate the agent negotiation mechanisms on the basis of process algebra. Moreover, private strategies in reaching agreements are discussed. Our approach reveals that reflections from other agents in negotiations can be captured and in turn provide sufficient information about how to manage the ontology along with proposed activities in ontology evolution.

1 Introduction

The vision of ontology for different purposes, such as Web-enabled applications, Web semantics, information systems, is receiving increasing attentions both from academia and industry. An ontology is an explicit *specification* of a shared *conceptualisation* [2]. The next generation of the WWW, the so-called Semantic Web, is based on using ontologies for annotating content with formal semantics. The benefits of using ontologies have been recognised in many areas such as knowledge and content management, e-commerce and Semantic Web. Approaches based on ontologies have shown the advantages for both information integration and system interoperability including reusability, extensibility, and verification analysis and so on. Based on the past experiences [4], however, it is unlikely to construct a monolith ontology due to the large scale, individual privacy, dynamics and heterogeneity. A typical example is virtual organisations (VOs). Interactions between agents on demand for any understanding is thus seen as a key technology at the knowledge view level. It is regarded as a feasible and effective way to find semantics in a certain business scenario at run-time.

Much work has been done in the field of ontology from the knowledge management perspective. Please refer to [3, 8] for thorough discussions. One distinguishing feature of those work is that in which essential operations have been designed to cater for management but without considering reflections during the

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1–7, 2005.

operating process. However, the trend of developing the Web and Web-based applications will inevitably continue, thus seeking a suitable approach as [5] that better reflects the reality of ontology management at run-time becomes one of the primary aspects of a successful ontology-supported application.

Agent technology [6, 10], embodied with autonomy, adaptation, and other features, fits well in this situation by providing a multi-agent system (MAS) environment with agents working together to solve problems that are beyond the overall capability of individual agents. By agents taking part in the ontology management at run-time, we expect that the bottleneck of information understanding on the basis of agents' interaction will be eased to some extent. The assumptions in this paper are that ontologies of different organisations are available and they are described by concepts and corresponding relations in a taxonomy. Any further manipulation is going on via interactions between agents (herein, an agent is used to annotate an actor/entity who plays particular roles in a process).

The rest of the paper is structured as follows. Section 2 presents an ontology definition. Section 3 details the ontology management on the basis of agent interactions by presenting a high level interaction diagram. Process algebra is used in modelling agent interaction and reaching an agreement. Section 4 studies negotiations with individual strategy profiles in VOs. Section 5 concludes our work and identifies potential further work.

2 Scenario and Ontology Definition

A concept definition, along with its relations and others if they exist, is more likely to be described in a taxonomy structure. In this paper, we follow the definition that the taxonomy tree which is a typical organisation to describe an ontology. We define an ontology which specifies a domain model, T, in terms of concepts and relations, as a tuple in the form of *(C, R, L, E)*, where C stands for a set of concepts including abstract concepts and primitive ones (sometimes called attributes) where abstract concepts include sub-concepts, while primitive ones are the baseline of people's knowledge of that domain; R defines binary relationships such as `is-a`, `part-of`, and `instance-of`; L is a set of logic operators such as "$\wedge$", "$\vee$" and "$\neg$"; E may be either "$=$" or "$\subseteq$" to specify relations between concepts. DeMorgan's laws are used to deduct expressions to only include conjunction and negation operators. We may simply notate ontology O as: $O = \bigcap_{i,j=1}^{m,n} R_j^i \cdot C_j^i$ under model T. $\forall$ concept C_j, $C_j := \bigcap_{j=1}^{n} R_j^i \cdot C_j^i$, where ":=" might be "$=$" or "$\subseteq$" according to T. Generally speaking, it is "$=$" except where "$\subseteq$" is specified. So we use formula $C_j = \bigcap_{j=1}^{n} R_j^i \cdot C_j^i$ in the following except where "'$\subseteq$" is specified. This formula implies that each concept might be decomposed into sub-concepts until primitive concepts in forms of C_j^i associated with relations R_j^i. It means that it can be described with the `part-of` relation in the end. In that way, all those sub-concepts are combined by operator "$\bigcap$" corresponding to the logical expressions in the conceptual model under a specific domain theory.

3 Ontology Management Based on Agent Interaction

Agents in a MAS may not share common goals, which more or less require run-time coordination of their activities and cooperation with others rather than hard-coded strategies at design-time. Bearing these in mind, we will investigate interaction protocols in process algebra and related issues in more detail in the following subsections.

3.1 Process Algebra as Interaction Model

An agent can perceive any change of environment [10] rather than being deaf-and-dumb during the operation. The interaction is likely to be described in a way of information flowing with certain restriction between interested agents. Two main types of actions are *sending* and *receiving* for a given *communication point*:

$\overline{a}(x).P$: meaning the action of sending the value along the point

$\underline{a}(x).P$: meaning the action of receiving the value along the point

Let $a, b, \ldots, z$ be a set of variables, information values, and interaction points, i.e. the `name` in process algebra; and P, Q, …, which are elements of process set $\mathcal{P}$. We assume that process set $\mathcal{P}$ represents the agents involved ranged over by P, Q, …, and the syntax of basic process algebra is defined as:

$$\begin{aligned} \mathcal{P} ::= \; & \alpha.P && \quad\quad \mathit{Prefix} \\ & P + Q && \quad\quad \mathit{Sum} \\ & P|Q && \quad\quad \mathit{Parallel} \\ & \mathbf{0} && \quad\quad \mathit{Nil} \\ & (\nu x)P && \quad\quad \mathit{Restricting} \end{aligned}$$

where α is the prefix of sending and receiving ranging over $\overline{a}(x)$ and $\underline{a}(x)$. Five basic constructors are prefix $\alpha.P$, *sum* $P{+}Q$, *parallel composition* $P|Q$, the empty agent $\mathbf{0}$, and restricting $(\nu x)P$, respectively. The purpose of describing reflections on each agent in a negotiation setting is achieved by defining the deduction rules that governing the negotiation process at run-time. Just as traditional process algebra [1], deduction operator "→" over agents like $P \rightarrow Q$ means that agent P can send a message, and agent Q can receive it on the basis of interactions between agents. In this respect, interactions between agents can be modeled by process algebra to highlight an agent's reflection and effects of interaction, and how it affects an agent's next action. By saying so, we attempt to describe agents' interactions by using process algebra. Before proceeding, in general, the interaction between two agents is defined as: $\overline{a}(x).P|\underline{a}(y).Q \rightarrow P|Q\{x/y\}$. Process $\overline{a}(x).P$ sends information x, through point a. After the information has been sent out, $\overline{a}(x).P$ becomes process P. P is the continuation of $\overline{a}(x).P$. In parallel, process $\underline{a}(y).Q$ receives that information through point a to transform to $Q\{x/y\}$, with information y substituted by information x via point a. The process algebra has its richness of mathematics. However, we are interested in the use of the notion for information flowing between agents in a particular negotiation setting.

Before we start to discuss this work, it is worth defining utility functions to tell agents how "good" the current result/state is. In terms of the result/state, as we assume that an agent is proactive, it should act in the direction of maximizing its welfare at each stage of a process, obviously, by considering environment changes. In this approach, we adopt a utility function defined in [10] as: $u : \mathit{\Omega} \rightarrow \Re$, where $\mathit{\Omega} = \{\omega_1, \omega_2, \ldots\}$ of states that agents have preferences over. For two elements ω and ω', if $u_i(\omega) > u_i(\omega')$, then we say ω is preferred by agent i at least as much as ω'. Every agent is willing to adopt a good strategy to maximize its expected utility. Clearly, optimal agent action $Ag_{optimal}$ which leads to the best performance under a certain circumstance is defined as:
$Ag_{optimal} = arg\ max_{Ag \in \mathcal{AG}} \sum_{\omega \in \mathit{\Omega}} u(\omega)$.

In addition to the interaction model on the basis of process algebra, how to solve conflicts or misunderstandings in a MAS is another issue that needs to be addressed. In the next, we will discuss the negotiation and negotiation strategies.

3.2 Negotiation Mechanisms

Negotiation in a MAS includes a negotiation set, a protocol, a collection of strategies, and rules [10]. Negotiation aims to eliminate conflicts/inconsistency to reach an agreement about specific negotiation issues. Obviously, efficient individual negotiation strategies in the correct direction of achieving the goal are at the heart at this stage. Normally a strategy is private, i.e., each agent has its own strategy profile which is invisible to other agents. In contrary to reaching an agreement for a price or shipping time, term negotiation here is an *open cry* setting to reach the agreement if conflicts in understanding each other occur. The negotiation strategy used is the strategic-negotiation model based on Rubinstern's model of alternative offers [9]. In the strategic negotiation model, a set of agents is defined as $\mathcal{AG} = \{Ag_1, Ag_2, \ldots\}$. It is assumed that the agents can take actions defined as $\mathcal{A}c = \{\alpha_1, \alpha_2, \ldots\}$, which are available from agents' action repertoire in a time sequence set $\mathcal{T} = \{0, 1, 2, \ldots\}$, that is known to the agents. A sequence set is defined as $\mathcal{S} = \{s_1, s_2, \ldots\}$, corresponding to the time sequence. $\forall t \in \mathcal{T}$ of a negotiation period, if negotiation is still going on without any deal reached, the agent, with its turn to make an offer at time t, will suggest a possible agreement with other agents which may choose one of the following three answers. Each of them may either `accept` an offer by choosing (`Yes`), `reject` an offer by choosing (`No`), or `opt out` an offer by choosing (`Opt`) of negotiation. If the offer is accepted by all agents, negotiation is terminated and then followed by implementation. Let $O = (C_{index}, t)$ be the offer that an agent makes at time period t, where C_{index} denotes the subgraph of a certain ontology hierarchy. Generally speaking, in order to reach an agreement in ontology management, simply choosing either `Yes`, `No`, or `Opt` is insufficient. For the purpose of providing more information in each negotiation round, the spectrum of the feedback state is extended to include the following activities:

- state 1: `No`. Implicitly, it may choose either one from the following three: (1) I really do not know; (2) I know something about it; (3) I know.
- state 2: `Agree to a certain degree`. it provides this feedback with an approximate percentage to enable the agent's opinion.
- state 3: `Yes`. It one hundred percent agrees with a specific negotiation issue suggested by the agent who is currently making an offer.
- state 4: `Opt out`. If at least one of the agents opts out of the negotiation, then the negotiation ends.

The process is shown in Figure 1, where each agent has its own strategy profile decided by its individual global view at certain period. On the left part of Figure 1, the interaction between agents is presented by process algebra like deduction rules. For simplicity, only two agents/processes (in a MAS we assume that the agent conducts the process, so we use them interchangeably) are shown. Generally speaking, interactions between agents can be depicted by parallel composition process $P|Q$ with $P, Q \in \mathcal{P}$. Any reflection from the previous or current round will affect an agent's next action. The negotiation process is shown on the right part of Figure 1.

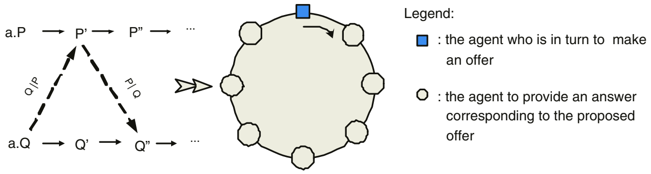

a.P
P'
P"
...
Q/P
P/Q
a.Q
Q'
Q"
...
Legend:
: the agent who is in turn to make an offer
: the agent to provide an answer corresponding to the proposed offer

Fig. 1. Negotiation process - strategy determined by interactions and agent's status

The process is thus described as a restricted token passing cycle with each agent making offers in turn, but other rational agents provide feedback (positive/negative) as much as possible. According to the above discussion, the offer proposed is defined as: $\mathcal{O} = (x, \mathrm{O})$, where x means the current agent's agreement upon a specific concept in the fashion of percentage. We are interested in reflections between agents at certain time period t. We will discuss agent interaction and its effect on the basis of process algebra next.

3.3 Agent Interaction

Misunderstanding is happening on the basis of both internal processes and external interactions. Process algebra is used to abstract the agent interactions in the face of exploration behaviours in MASs from a high level. In terms of the internal process, the grammar of agent P is defined as: $\alpha_1.P_1 + \alpha_2.P_2 + \ldots + \mathbf{0}$. We mentioned that agent P denotes its corresponding process as well. According to the syntax of basic process algebra, the process arrives its continuation after execution. We also consider that the agent reaches its continuation. Only when the process reaches $\mathbf{0}$, meaning no longer an activity to its environment, the process stops. On the contrary, the vision of agent's interactions is known as any agent P acting as sending a message, while another agent Q performing complementary action of receiving a message to continue its process. Thus the interaction is most likely be defined as: $\alpha.P|\alpha.Q$, where α is the prefix of sending/offering and receiving ranging over $\overline{a}(x)$ and $\underline{a}(x)$.

To the end, we have the following syntax rule of interactions in a negotiation system, where $Ag_i = (\alpha_1.P_1 + \alpha_2.P_2 + \ldots + \mathbf{0})$, and $i \in \mathbf{N}$.

$$Setting_{Neg} = Ag_1|Ag_2|\ldots$$

4 Strategy Profile in VO

(1) **Pre-VO Formation Agreement.** During this stage, from a loosely coupled net to becoming a closely coupled one, agents are strictly competitive within a boundary but cooperative across boundaries for the reason that each one wants to be successful in entering the next stage. By saying strictly competitive, we mean a scenario of zero-sum, namely, $\forall \omega \in \Omega, u_i(\omega) + u_j(\omega) = 0$.
(2) **During-VO Agreement.** Rather than strictly competitive, during this stage agents are cooperative and negotiation is in an environment of *open cry*. It is closely coupled at this stage. An agent will choose a "good" strategy to maximize its utility but no one has motivation to deviate and use another strategy because strategies used by agents are in *Nash equilibrium* [7]. Let $S^k_{i,j}$, a j^{th} suggestion at round i by the k^{th} agent, $A^k_{i,j}$, a j^{th} answer at round i by the k^{th} agent in percentage according to states defined in Subsection 3.2 with only exception of the `Opt out` occurrence. In term of *Nash equilibrium*, suppose ($S^{k_1}_{i,j}$, $S^{k_2}_{i,j}$, ...) are strategies in *Nash equilibrium* for agent k_1,k_2,... and so on, when an agent, for example k_1, chooses to play some other strategy, say $S^{k'_1}_{i,j}$, agent k_1 is likely to be worse than it would be if it plays $S^{k_1}_{i,j}$. By saying so, the feedback is therefore defined as $(A^k_{i,j} \cdot S^k_{i,j})$. Such kind of tracks will be kept by the agent who is in turn to make an offer at time period t. The negotiation terminates when all of agents' responses meet the defined criteria in negotiation setting. For example, in a VO where the main purpose is to catch the meaning of a concept, the rule to terminate the negotiation looks like:
R_1: every other agent reaches 90% or above agreement on a certain concept meaning.

Generally speaking, rational agents will not tell a lie if they really know answers of a specific negotiation issue or they know a little of it(in percentage). The negotiation proceeding to time period $t + 1$ is said that no agent has chosen `Opt out` but conflicts or misunderstandings still exist. Being aware of the reflection of other agents is critically important before any action for ontology management is taken. Although direct linkages are not explicitly stated in a VO, the sphere of direct influence is easy to be identified according to a particular business scenario. Moreover, tracks of other agents' responses are kept by the agent who offers an suggestion for a specific negotiation issue. There is no problem for this kind of information available in case it is needed.

(3) **Post-VO Ontology Management.** There is no more negotiation in solving misunderstanding of a particular concept in the post-VO stage. This stage is catering for ontology management. Suitable answers in forms of $(A_{i,j}^k \cdot S_{i,j}^k)$ for suggestion/offer $S_{i,j}^k$ could be selected based on the selection criteria of the agent and the feedback from others.

5 Conclusions and Future Work

We have proposed that interaction in the process algebra model symbolises the information flowing between agents and what impacts on agents. Each agent's strategy profile is determined by the feedback/environment and its current local global view. The novelty of our work lies in that we have highlighted the motivation of why ontology management is needed and how it adapts in accordance with the ontology evolving cycle. In addition, we have developed strategies to reflect the dynamic changes in it. Finally, as an key word in multi-agent systems, the interaction has been discussed throughly in both the mathematical and graphical aspects.

Although the approach in this paper is a promising way to run-time ontology management or even further to Semantic Web, some issues, such as evaluating the proposed negotiation strategies and developing relevant tools to support run-time ontology management, need to be further addressed.

References

1. Baeten, J. C. M., Weijland, W. P., *Process Algebra*, Cambridge University Press, ISBN 0521400430, 1990.
2. Gruber, T. R., Toward Principles for the Design of Ontologies Used for Knowledge Sharing, KSL-93-04, 1993. *http://ksl-web.stanford.edu/*.
3. Kalfoglou, Y., and Schorlemmer, M., Ontology Mapping: The State of the Art, *Knowledge Engineering Review*, 18(1), pp. 1-31, 2003.
4. Lenat, B. D. and Guha, V. R., *Building Large Knowledge-based Systems, Representation and Inference in the CYC project*, Addison-Wesley, Reading, Massachusetts, 1990.
5. Li, L., Yang, Y., and Wu, B., Agent-Based Approach towards Ontology Refinement in Virtual Enterprises, *Proc. of the 3rd International Conference on Active Media Technology* (*AMT 2005*), Japan, May 2005. To appear.
6. Luck, M., Mcburney, P., and Preist, C., A Manifesto for Agent Technology:Towards Next Generation Computing, *Autonomous Agents and Multi-Agent Systems*, 9(3), pp. 203-252, 2004.
7. Nash, F. J., Equilibrium Points in N-Person Games, *Proc. of NAS*, 1950.
8. *http://www.ontoweb.org/download/deliverables/D1.4-v1.0.pdf*
9. Rubinstein, A., Perfect Equilibrium in a Bargaining Model, *Econometrica*, 50(1), pp.97-109, 1982.
10. Wooldridge, M., *An Introduction to MultiAgent Systems*, John Wiley & Sons, ISBN 047149691X, 2002.

A Novel Approach for Developing Autonomous and Collaborative Agents

Nora Houari and Behrouz Homayoun Far

Department of Electrical and Computer Engineering, University of Calgary
2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4
{nhouari,far}@ucalgary.ca

Abstract. In this paper we present a novel approach that customize the BDI model to define a so-called "RBDIA: Rapport-Belief-Desire-Intention-Adaptation" as a generic method to support progress from individual autonomous agent concept towards a collaborative multiple agents. Rapport here refers to the component that connects an agent to its environment, whereas Adaptation module incorporates mechanisms of learning. The contribution of this paper is twofold: first, we develop a novel modeling approach that enable us to combine the internal and social structures of collaborative multigent, and second the proposed methodology is applied to a real-world application for assistance in product development process. We believe that the five proposed tiers for multi-agent systems (MAS) development serves for mastering the complexity and the difficulty of setting up effective autonomous collaborative MAS.

1 Introduction

The rapid development in computing is moving towards open, dynamic and ubiquitous environments in which devices, services, and software agents are all expected to seamlessly integrate and cooperate in support of human objectives, anticipating needs, negotiation for services, acting on users' behalf and delivering services in any where, and any time.

Software Agents have been recognized as a new paradigm to build complex software systems by simplifying the complexity of distributed systems and by overcoming the limitations of the existing computing approaches. They have been emerged as a next significant breakthrough in software development and a new revolution in software. Like human agent, a software agent can carry out a task, has its own characteristics, such as autonomous, social ability, reactivity, proactiveness, cooperative, learnable and adaptable. Therefore agent is considered as a natural abstraction of the real world, in which it can model the real world with its own goals and interacts with other agents to achieve mutual benefits. In this research work we embrace an integrated approach that captures both the design of the internal (mental) structures of single agents and the interactive shared (social) structures that underline multi agent collaboration. Here, we seek to depict the essential building components of collaborative multi-agents systems that adopt in one hand the belief desire intention (BDI) agent model which delineate the subjective world of the individual agents and in the

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 8–15, 2005.

other hand the common social (inter-subjective) world that constitute agents interaction.

In this paper we present a generic approach that customizes the BDI model to define a so-called "RBDIA: Rapport-Belief-Desire-Intention-Adaptation". Rapport here refers to the component that connects an agent to its environment, whereas Adaptation module incorporates mechanisms of learning.

The overall objective of the method is to be a solid foundation that serves for mastering the complexity of setting up effective multiagent systems that support progress from individual autonomous agent concept towards a collaborative multiple agents.

The BDI agent model was established in the mid nineteen eighties and has become the most known and studied model of practical reasoning agents that incorporate a prominent philosophical model of human reasoning and software agent implementation [8]. The beliefs, desires, intentions are the fundamental units of an agent based system designed for dynamic complex, and unpredictable environment. *Beliefs* here consist of a set of assertions about the state of the world which have an underlying propositional semantics. The general notion of assertion encompasses a full range of logical formalism [10]. *Desires* (or goals) are another key constituent of the system that represents some desired states. The intended plans or behaviors are *intentions*, which correspond to the third elementary component of the BDI model; intentions are simply executing states towards achieving the goals of the system.

Concerning the area of agent social interaction, numerous works on organizations modeling, roles and norms exist in the literature [2, 4]. Our approach differs from [3] in the way the BDI agent is captured. We follow the natural style of human thought by first capturing the desires, then intentions, and beliefs. Other work [7] shares with our research the way the mental structure of the agent is identified, however it varies in two aspects. First we follow the development process starting from the high level (architecture) and going down to the design and implementation. Second and most importantly we are incorporating in our methodology key concepts of intelligent agent, instead of only the mental (BDI model), we are adding the rapport that provide the social interaction of the agent, and adaptation component that is the ability of learning.

In our previous work [1] we proposed a novel agent interaction model in which a clear investigation of agents' interaction scenarios are provided with the appropriate reasoning techniques, building on that and adopting an integrated methodology, we are incorporating both the internal and the social interactive structures of MAS modeling.

The application use case of the proposed approach was applied to a major product development process at one of largest enterprise dealing with collaboration across multiple companies in the area of design and development of hardware components.

The remainder of this paper is structured as follows: in Section 2 we briefly give an overview of our approach of building collaborative multiagent systems. Section 3 presents the proposed RBDIA collaborative multiagent modeling. The application of the approach within a real world case application is described in section 4, and finally conclusion and future work are given in Section 5.

2 A New Approach for Modeling Collaborative Agents

Using an integrated approach of internal (mental) and interactive (social) structure we propose a robust agent model that extends the standard BDI agent model to incorporate the interaction with the environment that allow agent to manipulate itself to coordinate with others. The agents in the system can represent individuals (human or software) or collectives, including external stakeholders such as customers, regulators or suppliers, and internal entities such as staff, departments, or systems. From the perspective of the communicative action theory and building on the three-postulated world by Habermas [5] we are modeling our multiagent collaborative system from three perspectives:

- The *subjective* world, (how the agent perceives the world) that comprehends the feelings, beliefs, desires, experiences and intentions of the agent,
- The *common social* (inter-subjective) world that embodied the norms, commitments, agent relationships, and institutions to which the agents belong themselves, and which defines how agents stand towards each other, and,
- The *objective* world of objects and states of affairs (external world) that describes how things are.

Figure 1 depicts our new approach for modeling collaborative multiagent systems.

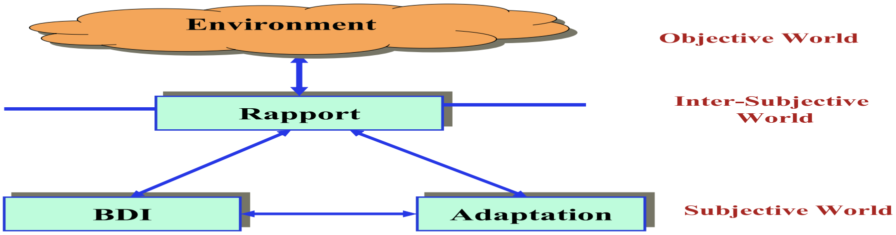


Fig. 1. The Proposed Method for Collaborative Multiagent Modeling

In our system development process, we are using the entire software development process and following a top down structured design method, taking into account system priorities and constraints, and ensuring that the system will achieve its objectives and requirements. After requirements acquisitions, we pursue the identification of the system architecture in terms of its components and connectors that help make the system more understandable, guide development implementation of the system and evaluation of the system for future modification. Here we process an iterative and incremental modeling approach, because of the limitation of the paper length, a detailed version of the work is provided separately on the authors' web site. In this paper we present only the analysis and design phases.

Our approach follows the natural style of human thought that capture *desires* (goals) in the beginning, under these goals it workout the correspond *intentions* (plans) in order to fulfill this goals, and collect the required information and knowledge, these are *beliefs*. To accomplish some of the plans agents need to communicate with their environment to perform some actions and or collect information this consists of *rapport* and at the same time they have to learn from their environment and manipulate themselves to collaborate with others, this is their ability to learn (*adaptation*). Figure 2 depicts our proposed approach.

In building our RBDIA model we first identify the desires from the system requirements these are the goals of the system, then capture the intentions theses are the plans, followed by the required knowledge which constitutes the beliefs both explicit and implicit in the form of potential deductions based on logical reasoning; to communicate with the environment agent need a dialog (rapport) established from communication language, protocol, norms, and contracts. As an agent, it needs to adapt, that is its ability to learn from its environment; machine leaning techniques are applied for building the learning capabilities. The building blocks of our agent modeling are shown in figure 3.

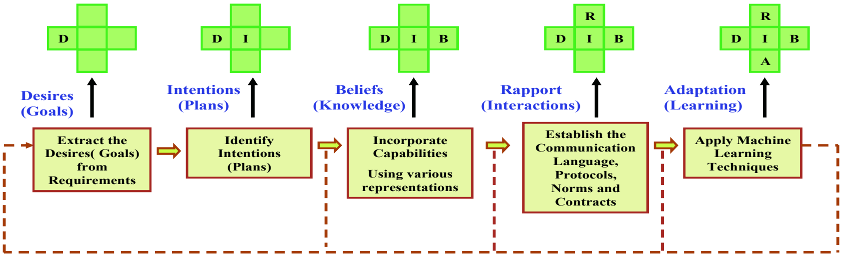

D
D I
D I B
R D I B
R D I B A
Desires (Goals)
Intentions (Plans)
Beliefs (Knowledge)
Rapport (Interactions)
Adaptation (Learning)
Extract the Desires(Goals) from Requirements
Identify Intentions (Plans)
Incorporate Capabilities
Using various representations
Establish the Communication Language, Protocols, Norms and Contracts
Apply Machine Learning Techniques

Fig. 2. A New Approach for Intelligent Collaborative Agents: RBDIA Model

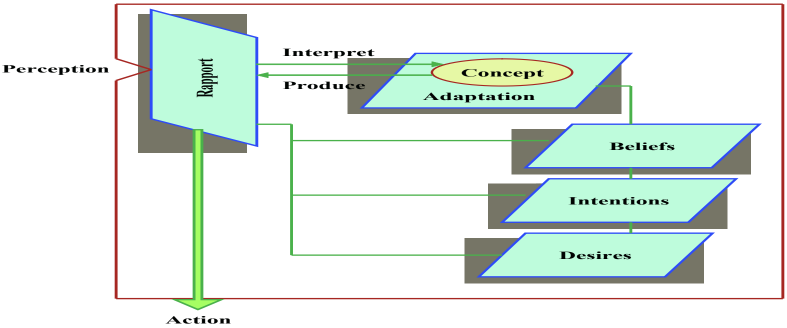


Fig. 3. A Generic Agent Building Blocks: RBDIA

3 Modeling Collaborative Agents

Our RBDIA modeling process follows the iterative and incremental software modeling phases; these are requirements, analysis, design, implementation, testing and maintenance. It has seven main steps: Problem determination and assessment, system architecture identification (components and connectors), desires (goals) identification, finding intentions (plans), designating the beliefs, specifying the rapport (dialog) with the environment, and finally building the learning ability (adaptation).

In this paper we underline only the requirements, analysis, and design phases. We used UML for the diagrams, propositional semantics for beliefs representation, and machine learning techniques for adaptation. These phases are briefly summarized in the following steps;

Step 1: Problem Determination and Assessment: This initial step describes the problem to be solved. It is the conceptualization of the required system from the customer's point of view, and it denotes the services and functionalities that the system has to provide.

Step 2: System Architecture: In this step we identify the system components and connectors. This step is a high level description of the system that helps make the system more understandable, intellectually manageable, guide development implementation and evolution of the system for future modification.

Step 3: Capturing the Desires (Goals): The desires (goals) of the system are specified from the requirements, they are abstract artifacts. We use UML use case, collaboration, activity and sequence diagrams to represent the system services more precisely from the external point of view, we identify what services the system should provide who are the actors, what they do and for what purpose without dealing with the internal details.

Step 4: Extracting the Intentions (Plans): From the external point of view, we create internal diagrams that show how the internal entities of the system interact, these help identify intentions for the previously specified goals and reveal more details that are not obvious to extract from the external point of view. Agent's plans (intentions) refer to sets of activities to be performed and here the roles are depicted (a role includes specification of what an agent is able to accomplish it range from a particular goal to a set of goals). The roles become agents when designed the beliefs to them.

Step 5: Identifying the Beliefs (Knowledge): Beliefs are the required knowledge that the plans (intentions) use to fulfill the goals (desires). These are stored in the form of a set of assertions, which comprise standards propositional operators (conjunctions, disjunctions, negations and application) a set of quantifiers (as in predicate calculus) modal and temporal operators, and other devices for quantifying assertions with a level of uncertainty, as well as ontological assertion. After identifying basic parts of system entities: agents, desires, intentions, and beliefs, we gave names to agents.

Step 6: Specifying the Rapport (Interaction): As agents need to collaborate to fulfill the requirements, they need to interact with one another and with the external environment. We use message passing as a unique form for method invocation. Agents distinguish different type of messages as speech acts and use a collaboration protocol, this protocol is a set of public rules that dictate the conduct of an agent with other agents to achieve a desired final outcome in sharing the knowledge and performing actions that satisfy the desired goals to fulfill some utility functions. Messages are in a defined XML (eXtensible Markup Language) format and transported using the Simple Object Access Protocol (SOAP).

Step 7: Agents Adaptation: Adaptation or ability to learn is one of the properties of an intelligent autonomous agent. In our approach we are investigating several machine learning techniques build in form of algorithms, For a neural network learning algorithm for example, we develop a procedure that includes, (i) the generation of knowledge for the agent (from step5), this knowledge is expressed in the form of IF… THEN rules, then (ii) the rules are used to train the neural network. (iii) The trained neural network is used in the situations that are not covered by the rules used in (ii).

At the end of the design phase we create the agent class diagram. It contains the agent name, desires, intentions, beliefs, rapport, and a learning algorithm. At this point the system is ready to implement. It is quite straightforward, because we have almost the whole ready to code agents' capabilities and behaviors. It is possible to go back to previous artifacts of the earlier steps and make some adjustments for consistency with the interaction and the learning ability of the agents.

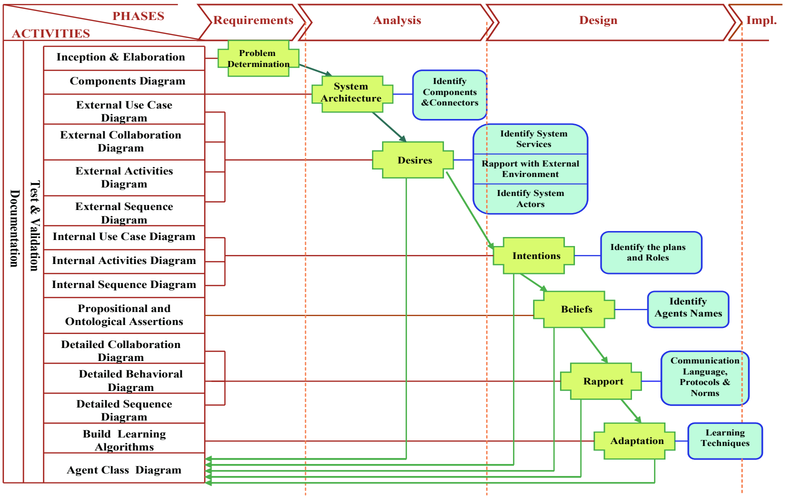

PHASES
ACTIVITIES
Requirements
Analysis
Design
Impl.
Documentation
Test & Validation
Inception & Elaboration
Components Diagram
External Use Case Diagram
External Collaboration Diagram
External Activities Diagram
External Sequence Diagram
Internal Use Case Diagram
Internal Activities Diagram
Internal Sequence Diagram
Propositional and Ontological Assertions
Detailed Collaboration Diagram
Detailed Behavioral Diagram
Detailed Sequence Diagram
Build Learning Algorithms
Agent Class Diagram
Problem Determination
System Architecture
Identify Components &Connectors
Desires
Identify System Services
Rapport with External Environment
Identify System Actors
Intentions
Identify the plans and Roles
Beliefs
Identify Agents Names
Rapport
Communication Language, Protocols & Norms
Adaptation
Learning Techniques

Fig. 4. RBDIA an Agent Development Process

4 Application Scenario of the Proposed Approach

The proposed approach was applied to a real world business context that specifically relate to a major manufacturing industry domain involves in hardware component design, and production. One of the real challenges of the manufacturing industry is how to shorten the time to go from a conceptual design to a product in a shorter period of time, also how to distribute the information of product usage to all appropriate entities, from depot maintenance, to supply units, to repair and manufacturing components and finally to original equipment manufacturers and aligned suppliers [9]. The solution to such complex problems, without time consuming and often-unavailable human intervention, can only be accomplished by invoking autonomous agents. Each agent specializes in small and distinct subsets of the overall objective of the system. Centralized control is not necessary; agents cooperate through their perception of the environment and learning ability. Their capabilities were delivered by embedding a few simple rules in each agent agenda. The rapport component of each agent helps agents to interact with each other and with a standard directory and legacy systems of the organization.

The platform chosen for the implementation of the system is Java Agent Development framework (JADE) [6]. We utilized the eXtensible Markup Language (XML) for agent communication, so when an agent sends message(s) to another agent (receiving agent), this message is in a defined XML format and is transported using the Simple Object Access Protocol (SOAP). The receiver agent parses the request message, processes its detail, and may return to the sender a replay also in XML format.

Our experience with the system indicates that the proposed solution is a flexible and constitutes an efficient tool with enormous capabilities for effective collaborative MAS.

5 Conclusions and Future Work

This paper presents a generic modeling approach that embrace both the internal (mental) structure of individual agent and the social (interactive) shared structure of multi-agent collaboration systems. Our approach follows the communicative action theory in modeling the collaborative multiagent system. We customized the BDI model to adopt the natural style of human thought that capture goals (desires) in the beginning, then workout the correspond plans (intentions) than collect the required knowledge (beliefs) to fulfill the goals. Agents are provided with interaction capabilities to coordinate with their environment (rapport) and endowed with learning algorithms (adaptation). We use an iterative and incremental software development method that follows a top down structured design for a software system applied to the design and development of engineering products development across multiple companies. Our approach serves for mastering the complexity of collaborative multiagent and the difficulty of setting up effective multiagent systems. Part of our ongoing work is investigating different learning algorithms. Future work will be links to other works in the area of social agent systems, especially in relation to the work of [2, 4, 10, 11].

References

1. Bahrouz, B.,”A Collective View and Methodologies for Software Agents Interaction". Transactions/Magazines/Conferences: IEEE Canadian Conference on Electrical and Computer Engineering (CCECE 2004), pp. 1249-1252, May 2004.
2. Dignum, V.:A model for Organizational Interaction: based on Agents, founded in Logic PhD Thesis,Tekst.,Proefschrift, Universiteit Utrecht, 2003.
3. Einhorn, M. Jeffery, A BDI Agent Software Development Process, MS Thesis, (Advisor: Chang-Hyun Jo), Department of Computer Science, University of North Dakota, May 2002.
4. Filipe, J., 2003. Information Fields in Organization Modeling using an EDA Multi-Agent Architecture. AAAI Spring Symposium 2003 on Agent-Mediated Knowledge Management (AMKM-2003). Stanford, USA.
5. Habermas, J. The Theory of Communicative Action: Reason and Rationalization of Society. Polity Press. Cambridge, 1984.
6. http://jade.tilab.com/
7. Jo, C.H., Guobin C., James, C., “A new approach to the BDI agent-based modeling” ACM Press New York, NY, USA, 2004.

8. Kinny, D., Georgeff, M., and Rao, A., A Methodology and Modeling Technique for Systems of BDI Agents, The 7th European Workshop on Modeling Autonomous Agents in a Multi-Agent World, LNAI, 61, 1996.
9. National Research Council, *Visionary Manufacturing Challenges for 2020*, National Academies Press,1998.
10. Sidney C. B.,Truszkowski, W., "Perspectives: An Analysis of Multiple Viewpoints in Agent-Based Systems" AMKM 2003: 368-387.
11. Vasconcelos, W., J.Sabater, C. Sierra and J. Querol, 2001. Skeleton-based agent development for electronic institutions. In proceedings of UKMAS.01.

The Effect of Alteration in Service Environments with Distributed Intelligent Agents

Dragan Jevtic[1], Marijan Kunstic[1], and Denis Ouzecki[2]

[1] University of Zagreb, Faculty of Electrical Engineering and Computing,
Unska 3, 10000 Zagreb, Croatia
{Dragan.Jevtic,Marijan.Kunstic}@fer.hr
http://www.tel.fer.hr/osoblje/index.hr.shtml
[2] Ericsson Nikola Tesla, Krapinska 45, 10000 Zagreb, Croatia
Denis.Ouzecki@ericsson.com

Abstract. The paper presents some properties of the communication network supported by the intelligent agents. The intelligent agents were placed into network nodes and they were immobile. They were used to regulate the transfer of mobile agents from the network input, through the routing nodes and finally, towards the service processing nodes. These nodes were the programs running on the computers. A new model of distributed and collaborating intelligent agents was designed and presented. Continuous adaptation and agent's collaboration was achieved by reinforcement *Q*-learning. For such a model the results show rapid tendency to reduce state time of mobile agents in the service. The obstructive effects expansion to the other agents when a change of processing capabilities in the region occurs was detected and described.

1 Introduction

Intelligent agents are being used for modeling rational but simple behaviors in wide range of distributed applications. They have received a variety of definitions. By general consensus, they must have some degree of autonomy, social ability, and combine pro-active and reactive behavior.

Real problems involve distributed open systems. An open system is one in which the structure of the system itself is capable of dynamic changing. The characteristics of such a system are that its devices are not known in advance; can change over time; and can consist of heterogeneous agents. These agents can be mobile, immobile and implemented for different services, processed at different locations, situations and environments. In an open environment, information sources, communication links and agents can appear and disappear unexpectedly. Their capabilities can change rapidly. Agent concepts, intelligent and mobile software agents have become a part of the system and service architecture of next-generation networks.

The intelligent agent represents an independent self-adaptive entity in the communication network [2, 3]. We have designed a heterogeneous agent system with three kinds of agents. We introduced intelligent agents (signed as selector agents - *SA*), client agents (*CA*), and service provider agents (*SPA*). *SA* and *SPA* agents were immobile, *CA* agents were assumed to be mobile. We assumed *CA* was entity acting on behalf of its user by moving through the network, but without intelligent capabilities. It migrated from node to node to find position and to request some processing from

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 16–22, 2005.

SPA. The task of *SA* agents was to regulate movement of *CA* agents through the network nodes towards the *SPA* agents. *SA* agents were distributed to the network nodes and collaborated locally on global routing in the network.

This paper is organized as follows: first, reinforcement learning is described and general description of Q-algorithm is specified in Section 2. The solution for *SA* agent distribution and collaboration in the network is introduced and described in Section 3. General model of topology is proposed. The inspected communication model is stated and described in Section 4. The results resenting the effects of the first adaptation and adoption to changes in *SPA* behavior are described in Section 5, followed by Conclusion.

2 Short Review of Reinforcement Learning

Reinforcement learning (RL) is an on-line training process. Training tries to follow environmental changes under specific assumptions. Thus, fundamental properties of RL are the automatic training that follows environmental changes with mathematically guaranteed convergence to optimal policy [3]. Agent reinforcement learning is a discrete-time system (t = 0, 1, 2, …). Assuming that the state and action spaces are finite and depend on the actions performed by an agent we get a typical finite Markov decision process. In this process a new state s_{t+1} after action a_t is determined with probability p:

$$p(s_{t+1}|s_t,a_t,s_{t-1},a_{t-1}, \ldots) = p(s_{t+1}|s_t,a_t). \tag{1}$$

Costs/rewards are accumulated during processing with a discount factor and can occur in certain states,

$$Q(s_o,a_o) = \lim_{N\to\infty} E[\sum_{t=0}^{N-1} \gamma^t g(s_t,a_t,s_{t+1}) \mid s_o,a_o], \tag{2}$$

where Q denotes the reward estimate for a starting pair state/action, s_0/a_0, and E denotes expectation. Subsequent actions a_t's are determined by an action policy, such as $a_t = argmax_a Q(s_t,a)$, g is reward, and γ is a discount factor. Q-learning algorithm utilize on-line updating, without explicitly use of probabilities [3]. Updating is based on the actual state transitions and is incremental,

$$\begin{aligned} Q_{t+1}(s_t,a_t) = (1-\alpha)Q_t(s_t,a_t) + \alpha[REWARD(s_t,a_t) \\ + \gamma \max_{a_{t+1}} Q_t(s_{t+1},a_{t+1})], \end{aligned} \tag{3}$$

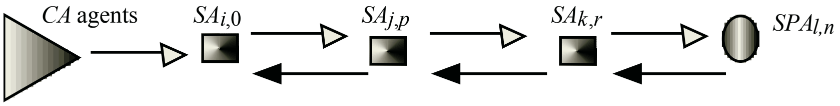
where α ∈ (0,1) is learning rate.

3 Reinforcement in Environment with Distributed Agents

A common characteristic of communication network is the presence of different kinds of nodes, interconnected by different kinds of links. Both nodes and the links have their own properties to be reached in order to perform the requested service. The agent's mission is to follow the network properties, discover fluctuations, and reach optimal solutions to the requests [1, 2]. In our model users' request were served by the mobile *CA* agents. Thus, *CA* agents had to move through the network. Mobility of *CA* agents controlled by *SA* agents. To support various configurations that might appear in the communication network, *SA* agents were organized in the layers and set to be immobile. Their task was to forward *CA* agents to the fastest output link. Their decisions depended on experience from previously completed actions. *SA* agent communication was thought to be simple for the following reasons:

- to avoid difficulties in case of mistakes and data loss - that might require complex messages,
- to minimize dependency of a message flow and communication network topology,
- to construct a self-adaptive agent system that can exist in the communication network.

That communication was limited only to the nearby *SA* agents and in a backward direction only. In short, *SA* agent forwarded *CA* agents from the input link towards next *SA* agent or to *SPA* agent depending on the concrete configuration (see example in the next section). *SA* agent determined its output link i.e. next node for *CA* observing attained *Q* values (see Tables in figure 1). *SA* agent having received *CA* agent immediately sent the reply with its predicted chosen route towards previous *SA* agent. Then he selected its output link for *CA* agent. The process went on until the *CA* agent had found *SPA* agent. *SPA* agent performed the requested task on behalf of *CA* agent. Task processing took some time and after completion, *SPA* sent the reply to *SA* agent. Based on this feedback information from its subsequent node, *SA* agent revitalized its knowledge.



Toward	$Q_i(0,p)$	Destination
$SA_{j,1}$	$Q_i(0,1)$	SPA_n
:	:	:
$SA_{j,m}$	$Q_i(0,m)$	SPA_n

Toward	$Q_j(p,k)$	Destination
$SA_{k,1}$	$Q_j(1,1)$	SPA_n
:	:	:
$SA_{k,h}$	$Q_j(p,h)$	SPA_n

Toward	$Q_k(k,r)$
SPA_1	$Q_k(k,1)$
:	:
SPA_n	$Q_k(k,n)$

Fig. 1. Generally, *SA* agents are located in corresponding network nodes. Corresponding prediction values is shown with tables

Agents are organized in a layered structure, with any number of *SA* agents in the layer (Figure 1). *CA* agent's movement is shown with white arrows. Movement of reply is denoted with black arrows. In our general structure we assumed one *SA* agent in *i*-th layer, where it used to collect *CA* agents. Zero, one or more *SA* agents could be placed in *j*-th layer, *k*-th layer and other such layers, if any, following in the same pattern. In this way each *SA* agent was responsible for its own small environment consisting of its input and output links. Final *l*-th layer consisted of *SPA* agents only. Since *SPA* agents performed final processing they could always be absolutely last (*l*-th) layer, independent of configuration of connections with preceding layers.

CA agents moved in the following way. The first *CA* agent came to 0-th node of *i*-th layer. Agent $SA_{i,0}$ chose some of their output links. General decisive factor was the first output link with fewer *Q* value (see Table with prediction values in figure 1). If previous experience were missing, it would be a random choice.

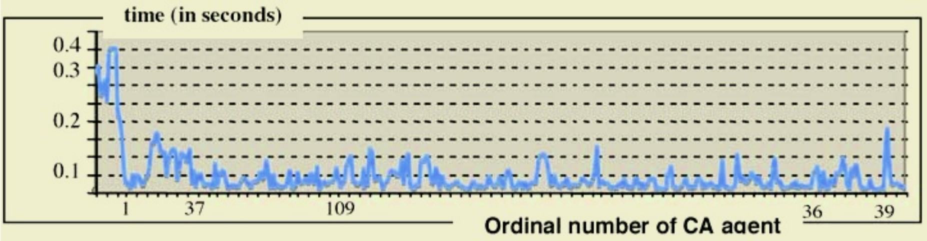


Fig. 2. The time of *CA* agents spent to perform the service. Transporting toward and processing by *SPA* are included

Then, *CA* agent moved through the chosen link towards *p*-th *SA* agent in the next *j*-th layer. By receiving *CA*, *p*-th *SA* agent immediately sent the reply to the 0-th node. The reply sent back, included *p*-th agent's own general prediction of all its output links based on previous experiences. At that moment $SA_{j,p}$ agent had to chose its output link towards the *k*-th layer, and so on, if there are more layers. When *CA* agent came to the last *SA* layer (*k*-th in figure 1), *r*-th *SA* agent had to select one of the present *SPA* from *l*-th layer. The reply from *SPA* to *SA* arrived after completion of *CA* request processing. Waiting time for the reply was a key parameter used by *SA* in *k*-th layer to determine update measure for the *Q*-value (see prediction values for $SA_{k,r}$). Each new *CA* agent initiated the actions in *SA* layers asynchronously from *i*-th towards *l*-th. If this process goes on, the system converges to the state of minimal staying time for *CA* requests in transfer and processing. Reduced staying time of *CA* requests was due to finding out optimal paths and optimal *SPA* agents. A typical example for the system configured in figure 3 is shown in figure 2. By knowing accumulation based on the first 18 agents, the system significantly reduces staying time for all subsequent *CA* agents.

4 Route Selection Model Based on Intelligent Agents

The network model supported by intelligent agents was investigated. Exactly two intelligent agents were used to reach optimal route for mobile *CA* agents (figure 3). We had two paths - one with two layers of decision (through SA_1 and SA_2) and one with one layer of decision (through SA_1).

SA_1 corresponded to the *i*-th layer and SA_2 to the *j*-th layer of the model from figure 1. Agent SA_1 had to decide whether to send *CA* to the SPA_1 or to pass a decision to the agent SA_2. Agent SA_2 had to make decisions by observing behavior of the group comprising of agents SPA_2 and SPA_3. We assumed that *CA* agents not have returned to their users via same nodes. Therefore, when *CA* was successfully served by *SPA*, the reply was sent to *SA* agent who determined the link towards *SPA* agents. Because of different factors in their environment, *SPA* agents can change their properties.

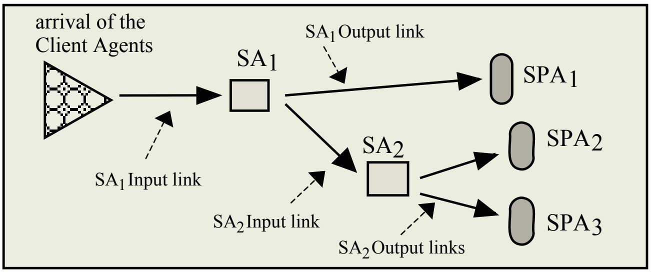


Fig. 3. A service system model in which two intelligent *SA* agents configured the paths for *CA* agents towards *SPA* agents

To examine the affects that might be essential for *SA* decisions we had experimented in real time with *CA*, *SA* and *SPA* agents distributed in different computers, interconnected through our local area network. We assumed routing to be 10 times faster than *CA* request processing. To make decision *SA* agents had to receive replies when *CA* agents completed their requests by *SPA*. For these actions *SA* agents brought their decisions faster when:

- arrival of *CA* agents was much more slower then the total time of their transferring from the *SA* to *SPA*, processing by the *SPA* and reply,
- arrival of the reply to *SA* agent succeeded at least before the arrival of next *CA* agent that had to be routed,
- there were no gaps in responses from *SPA* or/and *SA* agents.

A typical traffic situation was selected and applied to the model with two *SA* agents from figure 3. Behavioral change was applied instantly to the agents SPA_2 and SPA_3, by exchanging their properties. This alteration effect was realized by replacing SA_2 output links. That was done after obtaining stable and low standing time for *CA* agents and for *SPA* request processing in the system by training with 400 *CA* agents. Because both agents belonged to the same group, the group property remained unchanged. Before discovering the change in its group, agent SA_2 continued with routing *CA* agents employing the acquired knowledge which was truthful, but only within the scope of that change. In practice, this situation can occur whenever re-routing of traffic is applicable between two or more destinations. But in the environments loaded with intelligent agents organized in this way, the change in the group was immediately reflected on previous layers of a decision.

5 Results

The following diagrams show time difference between time of *CA* agent leaving *SA* agent and the reply to *SA* agent. Upper curve denotes arrival of the reply, and lower curve indicates departure time of *CA* agent. The numbers on *x*-axis are associated with the order of agent's arrival. For example, number 7 on *x*-axis denotes 7-th *CA* agent that arrived to that node. The numbers on *y*-axis denote actual computer time. In reality, this time expressed in seconds covers the monitoring period.

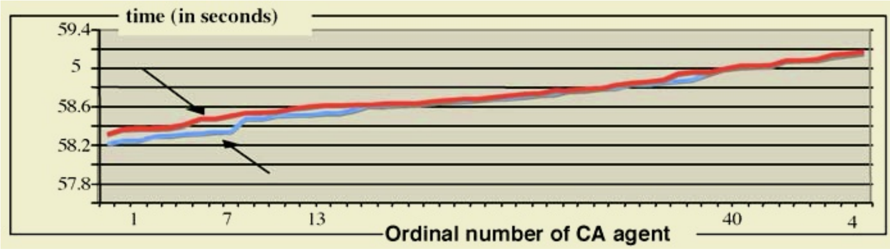


Fig. 4. Agent SA_1 started working without previous experience. Time difference after 18 *CA* agents markedly decreased. The curves are closest to each other

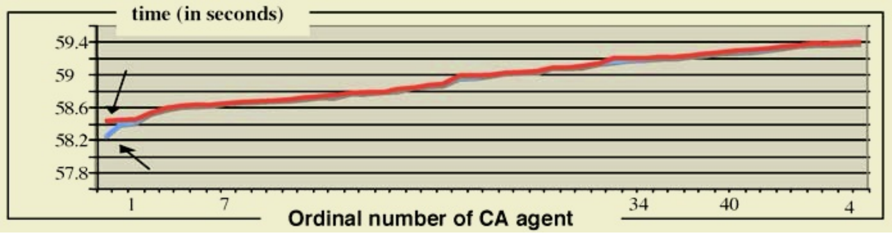


Fig. 5. Agent SA_2 started working without previous experience. Time difference after 4 *CA* agents markedly decreased. Learning period shows faster adaptation than for agent SA_1

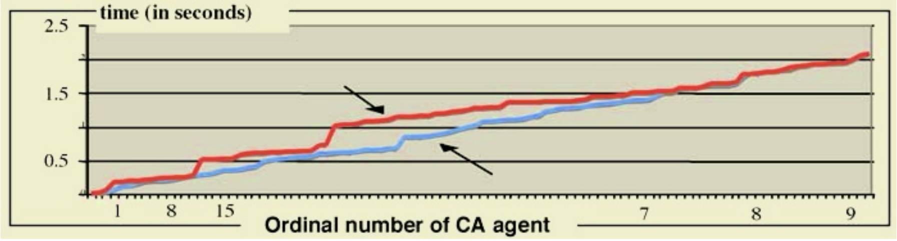


Fig. 6. Agent SA_1 routed *CA* agents to SPA_1 more frequently than towards SA_2. That was the consequence of predictions received from SA_2

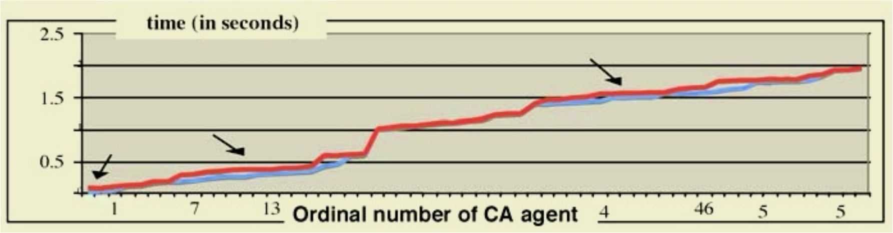


Fig. 7. Agent SA_2 routed *CA* agents towards previously faster *SPA*. It has to discover the change by exploration. During adaptation, SA_2 decreased its predictions about capabilities of its group

First two diagrams (figures 4 and 5) show rapid self-adaptation of *SA* agents, manifested in minimization of time for the receipt of both *SA* agents' reply. This means that they routed *CA* agents towards the fastest available *SPA* agents. Next two diagrams (figures 6 and 7) indicate possible significant decrease in system performance because of instantly replacing behaviors between *SPA* agents controlled by the same *SA* agent, i.e. in one subgroup of the system.

The learning parameters are the same for both *SA* agents, learning rate was set to 0.88 and exploration was set to probability 10%. New agents arrivals rate was modeled by exponential distribution with mean arrival rate of 50 ms per *CA* agent. Request holding time by SPA agents was exponentially distributed with mean holding duration approximately 60, 20 and 80 ms for SPA_1, SPA_2 and SPA_3, respectively.

6 Conclusion

Communication network model built by the intelligent agents was created. Cooperation between intelligent agents was a crucial function for successful routing of mobile agents through the network nodes. This cooperation was initially on the local level, among adjacent agents. The agents were placed in a computer network environment and tested in real time. The outcomes of measurements show a high degree of adaptability for a variety of traffic states. The situation existing when instant change in the behavior of the group elements appeared was recognized as the cause that can temporally change decisive factor of intelligent agents.

References

1. Jevtic D., Kunstic M., Jerkovic N.: The Intelligent Agent-Based Control of Service Processing Capacity. Lecture Notes in Computer Science/ LNAI 2774, Part 2. Springer-Verlag, Berlin Heidelberg New York (2003) 668-674
2. Bertsekas, D. P., Tsitsiklis, J. N.: Neuro-Dynamic Programming. Athena Scientific, MIT, Belmont, Massachusetts (1996)
3. Watkins, C.J.C.H., Dayan, P.: Q-learning. Machine Learning, Vol. 8, Kluwer Academic Publishers, (1992) 55-68

Managing Collaboration in a Multiagent System

John Debenham and Simeon Simoff

Faculty of IT, University of Technology, Sydney, NSW, Australia
{debenham,simeon}@it.uts.edu.au

Abstract. In a multiagent system agents negotiate with one another to distribute the work in an attempt to balance the incompatible goals of optimising the quality of the result, optimising system performance, maximising payoff, providing opportunities for poor performers to improve and balancing workload. This distribution of work is achieved by the delegation of responsibility for sub-processes by one agent to another. This leads to estimates of the probability that one agent is a better choice than another. The probability of delegating responsibility to an agent is then expressed as a function of these probability estimates. This apparently convoluted probabilistic method is easy to compute and gives good results in process management applications even when successive payoff measurements are unpredictably varied.

1 Introduction

Process management is an established application area for intelligent systems [1] [2]. Rule-based systems can manage production workflows, and multiagent systems are well-suited to manage the more complex goal-driven and emergent (or, knowledge-driven) processes. A multiagent system [3] manages the range of business processes from production workflows to high-level emergent processes. In that system each agent represents the interests of either a human player or a system resource such as a corporate database.

The way in which a process management system attempts to address corporate priorities may lead to conflicting goals. Here, agents negotiate with one another to distribute the work by delegating responsibility for a sub-process to another agent. Negotiation is conducted using contract nets with focussed addressing [4]. An agent who wishes to delegate responsibility announces the sub-process and any constraints on it. Other agents respond with bids for the work that are composed of values for a predetermined set of payoff measures. The receiving agent then evaluates these bids in the light of its own experience in dealing with each agent. Each agent is responsible for bringing the sub-processes for which it is responsible to a satisfactory conclusion within the agreed sub-process constraints [5]. This delegated responsibility may be further delegated — at least in part — to other agents. The *responsibility delegation problem* belongs to the class of resource allocation games which are inspired by the 'El Farol Bar' problem — [6] describes recent work in this area.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 23–29, 2005.

2 Delegation

A delegation strategy is a mechanism for deciding who to give responsibility to for doing what. If agent X_0 wishes to delegate responsibility for part of a process then it does so in three steps. First X_0 announces a proposal to a focussed subset of n agents in its community $\{X_1, \ldots, X_n\}$ who are *candidates* for the delegation. Second X_0 receives bids from these n agents. Third X_0 chooses an agent from this set. So agent X_0's delegation strategy will simply determine which of these n agents is to be the "lucky one". The strategies considered here achieve this indirectly by determining instead n probabilities $\{\mathbf{P}_1, \ldots, \mathbf{P}_n\}$ where $\mathbf{P}_i$ is the probability that the i'th agent will be selected, and $\sum_i \mathbf{P}_i = 1$. The choice of the agent to delegate to is then made with these probabilities. By expressing the delegation strategy in terms of probabilities, the agents have the flexibility to balance conflicting goals, such as achieving process quality and process efficiency.

$\mathbf{P}(X_i \gg)$ denotes the *rank* of agent X_i. Rank is "the probability that agent X_i is the 'best' choice of agent, chosen from $\{X_1, \ldots, X_n\}$, to delegate responsibility to". A *delegation strategy* is a set $\{\mathbf{P}_1, \ldots, \mathbf{P}_n\}$ where $\sum_i \mathbf{P}_i = 1$. The delegation strategies described here are determined by: $\mathbf{P}_i = f(\mathbf{P}(X_i \gg))$ for some function f that preserves the constraint $\sum_i \mathbf{P}_i = 1$. So given that there is some means of determining the $\mathbf{P}(X_i \gg)$ the question of which is the 'best' delegation strategy reduces to which is the best choice of function f. The choice of function f is in addition to, and independent of, the way in which agent X_0 chooses to specify $\mathbf{P}(X_i \gg)$.

The probabilities $\mathbf{P}(X_i \gg)$ are calculated at the time at which the delegation is made. The basis for determining these probabilities includes:

- objective (historic) measurements of prior performance
- estimates of future performance
- subjective estimates

for individual agents, and for the effectiveness of the interactions in particular groups of agents. We assume that for each of the n agents, these measurements and estimates can be combined to give a single *expected payoff* vector for each agent $\underline{\nu}_i$. There is no need to be over prescriptive on the form of this vector — in the examples discussed below the vector is a pair of numbers representing the expected mean and expected standard deviation of the payoff for each agent. We do assume however that the payoff vector contains sufficient information to estimate the probability that one agent is expected to deliver higher payoff than another in some sense by the agent X_0.

Delegation strategies used by humans can be quite elementary; delegation is a job which some humans are not very good at. A delegation strategy attempts to balance conflicting principles including: maximising payoff, maximising opportunities for poor performers to improve, and balancing workload. Payoff is some combination of: the expected value added to the process [7], the expected time and/or cost to deal with the process, and the expected likelihood of the process leading to a satisfactory conclusion [8], whilst satisfying the process constraints. The next section discusses the sorts of payoff measures and estimates that are available, and that are combined to give a value for the expected payoff vector $\underline{\nu}_i$ for each agent. These measurements are then used by agent X_0 to determine $\mathbf{P}(X_i \gg)$, and then in turn to determine the delegation strategy $\{\mathbf{P}_i\}$.

3 Payoff: $\{\nu_i\}$

Agent X_0 continually measures the performance of itself and of other agents in the system. The agents in the process management system are based on a hybrid BDI architecture [9] and are built in Jack. For each goal that an agent is committed to, the agent must choose a plan to achieve it. These plans may involve delegation. The payoff measures together form one class of input to the process for estimating the set of ranks $\{\mathbf{P}(X_i \gg)\}$ which is described in Sec. 4.

There are five measures for agent X_0. Three are: *time*, *cost* and *likelihood of success* which are attached to all of its plans and sub-plans. The remaining two are *value* and *delegate* parameters that are attached to other agents. Time is the total time taken to termination. Cost is the actual cost of the of resources allocated. For example, if an agent has a virtual document in its ‘in-tray’ then the time observation will be the total time that that document spent with that agent, and the cost is derived from the time that the agent — possibly with a human ‘assistant’ – actually spent working on that document[1].

The three measures time, cost and likelihood of success are recorded every time a plan or delegation is activated for a goal. This generates a large amount of data whose significance can reasonably be expected to degrade over time. Rather than record the raw data it is summarised using the geometric mean. Given a set of observations $\{ob_i\}$ where ob_1 is the most recent observation:

$$\frac{\sum_{i=1}^{n} \alpha^{i-1} \times ob_i}{\sum_{i=1}^{n} \alpha^{i-1}}$$

is the geometric mean where α is some constant, $0 < \alpha < 1$. If $\alpha = 0.85$ then “everything more than twenty trials ago” contributes less than 5% to the geometric mean. If $\alpha = 0.70$ then “everything more than ten trials ago” contributes less than 5% to the geometric mean. This method of summarising the observations does not take into account the time at which the observations were made, or the amount of time that has elapsed between observations. If this is an issue then the definition of the geometric mean is modified. If the observations $\{ob_i\}$ for some parameter p are drawn from a symmetrically distributed population then the geometric mean gives a point estimate of the mean of the population μ_p. The geometric mean is preferred to the (conventional) mean because by paying more attention to recent observations it adapts to fundamental changes in the population. Further, if these observations are drawn from a normal population then:

$$\frac{\sum_{i=1}^{n} \alpha^{i-1} \times \mid ob_i - \mu_p \mid}{\sum_{i=1}^{n} \alpha^{i-1}}$$

is a (geometric) estimate of $\sqrt{\frac{2}{\pi}}$ times the standard deviation of parameter p, σ_p. Where the constant α is determined empirically as described above. We now assume that the parameters time and cost are normally distributed. Theoretically this is a radical assumption. Practically it is “not unreasonable”, and is highly desirable because the geometric means are updated with the simple formulae:

[1] Cost here does not refer to costs incurred by the plan

$\mu_{p_{new}} = (1-\alpha) \times ob_i + \alpha \times \mu_{p_{old}}$
$\sigma_{p_{new}} = (1-\alpha) \times \mid ob_i - \mu_{p_{old}} \mid + \alpha \times \sigma_{p_{old}}$
with starting values $\mu_{p_{initial}}$ and $\sigma_{p_{initial}}$. The likelihood of success observations are binary – ie "success" or "fail" – and so the likelihood of success parameter is binomially distributed, which is approximately normally distributed under the standard conditions.

Finally, consider measurements of the *delegate* parameter for each agent. This parameter is a pair of related parameters. First, w_i^{in} is the amount of work delegated *to* agent i in a given discrete time period. Second, w_i^{out} is the amount of work delegated *by* agent i in the same discrete time period. In a similar way to time and cost, the mean delegate estimate for agent X_i is made using
$delegate_{new} = (1-\alpha) \times w_i + \alpha \times delegate_{old}$,
where w_i is the most recent observation (either delegations 'in' or delegations 'out') for agent X_i. In this formula the weighting factor α is chosen on the basis of the number of individuals in the system, and the relationships between the length of the discrete time interval and the expected length of time to deal with the work. The two components of the *delegate* parameter do not represent workload. For example, if responsibility is delegated and then re-delegated, the delegate in estimate of the first individual is not reduced. The *delegate* parameter is used by delegation strategies that address the frequency with which agents accept responsibility for doing things, and the frequency with which agents ask for assistance. The two components of the *delegate* parameter are not normally distributed and the standard deviation is not estimated. The *delegate* and *value* estimates are associated with individuals. The *time*, *cost* and *likelihood of success* estimates are attached to plans and delegations.

The three parameters *time*, *cost* and *likelihood of success* are assumed to be normally distributed subject to "all things being equal". One virtue of the assumption of normality is that it provides a basis on which to query unexpected observations. Having made observation ob_{i+1} for parameter p, estimates for μ_p and σ_p are calculated. Then the next observation, ob_i, should lie in the confidence interval: $(\mu_p \pm \alpha \times \sigma_p)$ to some chosen degree of certainty. The set of observations $\{ob_i\}$ can progressively but gradually change without individual observations lying outside this confidence interval; for example, an individual may be gradually getting better at doing things. But if an observation lies outside this confidence interval then there is grounds, to the chosen degree of certainty, to ask why it is outside.

The measurement ob_i may lie outside the confidence interval for various reasons. If the reason Γ is to be taken into account then some forecast of the future effect of Γ is required. If such a forecast effect can be quantified perhaps by simply asking an agent — then the perturbed values of $\{ob_i\}$ are corrected to $\{ob_i \mid \Gamma\}$ otherwise single perturbed values are ignored.

4 Rank: $\{\mathrm{P}(X_i \gg)\}$

Delegation of responsibility is achieved through *negotiation* during which agents are invited to express an interest in doing something. This negotiation uses contract nets with focussed addressing [4]. The use of a multi-agent system to manage processes expands the range of feasible delegation strategies.

A *bid* consists of the five pairs of real numbers (*Constraint*, *Delegate*, *Success*, *Cost*, *Time*). The pair *Constraint* is an estimate of the earliest time that the agent could address the taskie ignoring other non-urgent things to be done, and an estimate of the time that the agent would normally address the task if it "took its place in the queue". The pair *Delegate* is estimates for total delegations "in" and total delegations "out". The pairs *Success*, *Cost* and *Time* are estimates of the mean and standard deviation of the corresponding parameters as described above. Each bidding agent is assumed to be honest in reporting these constraint, delegate, success, time and cost estimates. The receiving agent X_0 then:

- attaches a subjective view of the Value of each bidding agent to that agent's bid;
- assesses the extent to which a bid should be downgraded — or not considered at all — because it violates or threatens process constraints;
- evaluates each bid — which may have been downgraded — by applying a function g_0 to (*Value*, (*Constraint*, *Delegate*, *Success*, *Cost*, *Time*)), and
- calculates the rank $\mathbf{P}(X_i \gg)$ for each of the bidding agents.

If there are no acceptable bids then the receiving agent "thinks again".

The method described above estimates the probability that one agent is a better choice than another. It may be extended to estimate the probability that one agent is a better choice than a number of other agents. For example, if there are three agents to choose from, A, B, and C, then:
$\mathbf{P}(A \gg) = \mathbf{P}((A \gg B) \wedge (A \gg C)) = \mathbf{P}(A \gg B) \times \mathbf{P}((A \gg C) \mid (A \gg B))$
The difficulty with this expression is that there is no direct way of estimating the second, conditional probability. This expression shows that:
$\mathbf{P}(A \gg B) \times \mathbf{P}(A \gg C) \leq \mathbf{P}(A \gg) \leq \mathbf{P}(A \gg B)$
By considering the same expression with B and C interchanged:
$\mathbf{P}(A \gg B) \times \mathbf{P}(A \gg C) \leq \mathbf{P}(A \gg)$
$\mathbf{P}(A \gg) \leq \min[\mathbf{P}(A \gg B), \mathbf{P}(A \gg C)]$
So for some $\tau_A \in [0,1]$:
$$\mathbf{P}(A \gg) = \mathbf{P}(A \gg B) \times \mathbf{P}(A \gg C) + \tau_A \times [\min[\mathbf{P}(A \gg B), \mathbf{P}(A \gg C)] - \mathbf{P}(A \gg B) \times \mathbf{P}(A \gg C)]$$
Similar expressions may be constructed for the probabilities that B and C are the best agents respectively. This is as far as probability theory can go without making some assumptions. To proceed assume that: $\tau_A = \tau_B = \tau_C = \tau$; this assumption is unlikely to be valid, but it should not be "too far" from reality. Either A or B or C will be the best plan, so the sum of the three expressions for the probabilities of A, B and C being the "best" plan will be unity. Hence:
$\tau = \frac{1-d}{q-d}$ where:
$$d = [(\mathbf{P}(A \gg B) \times \mathbf{P}(A \gg C)) + (\mathbf{P}(B \gg C) \times \mathbf{P}(B \gg A)) + (\mathbf{P}(C \gg A) \times \mathbf{P}(C \gg B))]$$
$$q = [\min[\mathbf{P}(A \gg B), \mathbf{P}(A \gg C)] + \min[\mathbf{P}(B \gg C), \mathbf{P}(B \gg A)] + \min[\mathbf{P}(C \gg A), \mathbf{P}(C \gg B)]]$$
This expression for τ is messy but is easy to calculate. The probability that each of the three agents A, B and C is the "best" choice is $\mathbf{P}(A \gg)$, $\mathbf{P}(B \gg)$ and $\mathbf{P}(C \gg)$.

5 Strategy: $\{\mathrm{P}_i\}$

Each agent determines its own delegation strategy that it uses to evaluate bids and manage the delegation process. This section describes various functions f as defined in Sec. 2. In doing this the agent has considerable flexibility first in defining payoff, determining rank and in specifying the strategy itself.

Given a sub-process, suppose that we have some expectation of the payoff ν_i as a result of choosing the i'th agent from the set of candidates $\{X_1, \ldots, X_n\}$ to take responsibility for it, and of the probability $\mathbf{P}(X_i \gg)$ that X_i is the best choice. A *delegation strategy* at any given time is a set $S = \{\mathbf{P}_1, \ldots, \mathbf{P}_n\}$ where $\mathbf{P}_i$ is the probability of delegating responsibility at that time for a given task to agent X_i chosen from $\{X_1, \ldots, X_n\}$. The method then selects an agent/task pair stochastically using the delegation strategy.

Four delegation strategies are described. If agents' community culture is to choose the agent whose expected payoff is maximal then the delegation strategy *best* is:

$$\mathbf{P}_i = \begin{cases} \frac{1}{m} & \textit{if}\, X_i \text{ is such that } \mathbf{P}(X_i \gg) \text{ is maximal} \\ 0 & \textit{otherwise} \end{cases}$$

where $\mathbf{P}(X_i \gg)$ means "the probability that X_i will have the highest payoff" and m is such that there are m agents for whom $\mathbf{P}(X_i \gg)$ is maximal. In the absence of any other complications, the strategy best attempts to maximise expected payoff. Using this strategy, an agent who performs poorly may never get work. Another strategy *prob* also favours high payoff but gives all agents a chance, sooner or later, and is defined by $\mathbf{P}_i = \mathbf{P}(X_i \gg)$. The strategies *best* and *prob* have the feature of 'rewarding' quality work (ie. high payoff) with more work. If community culture dictates that agents should be treated equally but at random then the delegation strategy *random* is $\mathbf{P}_i = \frac{1}{n}\ \forall i$.

The strategy *random* completely ignores the bids. If the community culture dictates that each task should be allocated to m agents in strict rotation then the delegation strategy *circulate* is:

$$\mathbf{P}_i = \begin{cases} 1 & \textit{if} \text{ this is the j'th trial } and\ i \equiv j \ (\text{modulo } n) \\ 0 & \textit{otherwise} \end{cases}$$

The strategies random and circulate attempt to balance workload and ignore expected payoff. The strategy circulate only has meaning in a fixed population, and so has limited use.

A practical strategy that attempts to balance maximising "expected payoff for the next delegation" with "improving available skills in the long term" could be constructed if there was a model for the expected improvement in skills — ie a model for the rate at which agents learn. This is not considered here. An admissible delegation strategy has the properties:

if $\mathbf{P}(X_i \gg) \ > \ \mathbf{P}(X_j \gg)$ *then* $\mathbf{P}_i > \mathbf{P}_j$
if $\mathbf{P}(X_i \gg) \ = \ \mathbf{P}(X_j \gg)$ *then* $\mathbf{P}_i = \mathbf{P}_j$
$\mathbf{P}_i > 0 (\forall i)$ and $\sum_i \mathbf{P}_i = 1$
So the three strategies *best*, *random* and *circulate* are not admissible.

An admissible strategy will delegate more responsibility to agents with a high probability of having the highest payoff than to agents with a low probability. Also with an admissible strategy each agent considered has some chance of being given responsibility. The strategy prob is admissible and is used in the system described here.

To generalise the above, suppose that an agent selects a strategy from the infinite variety of admissible strategies: $S = \delta \times \mathit{best} + \epsilon \times \mathit{prob} + \phi \times \mathit{random} + \gamma \times \mathit{circulate}$ will be admissible if $\delta, \epsilon, \phi, \gamma \in [0, 1]$, $\delta + \epsilon + \phi + \gamma = 1$ and if $\epsilon > 0$. This leads to the question of how to select a strategy. As *circulate* is only meaningful in stable populations it is not considered here. Real experiments to evaluate delegation strategies are just not viable. Laboratory simulation experiments are cheap and indicate how the strategies should perform.

Using three basic built-in strategies, the agent then specifies a delegation strategy for the chosen definition of payoff. In this way the agents handle sub-process delegation automatically. The system has been trialed on applications in a university administrative context and in an eMarket. Three delegation strategies $[\delta = 0.5, \epsilon = 0.5, \phi = 0]$, *prob* and $[\delta = 0, \epsilon = 0.5, \phi = 0.5]$ represent varying degrees of the "aggressive pursuit of payoff" and have been declared "reasonable" in limited trials. The method is easy to compute and gives good results in process management applications even when successive payoff measurements are unpredictably varied.

References

1. Huhns, M.N. and Singh, M.P., Managing heterogeneous transaction workflows with cooperating agents, In N.R. Jennings and M. Wooldridge, (eds), Agent Technology: Foundations, Applications and Markets, Springer-Verlag: Berlin, Germany, (1998), pp. 219–239.
2. Jennings, N.R., Faratin, P., Norman, T.J., O'Brien, P. and Odgers, B., Autonomous Agents for Business Process Management, Int. Journal of Applied Artificial Intelligence 14 (2) 145-189 (2000).
3. Debenham, JK., Managing e-Market Negotiation in Context with a Multiagent System, in proceedings Twenty First International Conference on Knowledge Based Systems and Applied Artificial Intelligence, ES'2002: Applications and Innovations in Expert Systems X, Cambridge UK, December (2002).
4. Durfee, E.H., Distributed Problem Solving and Planning, in Weiss, G. (ed), Multi-Agent Systems, The MIT Press, Cambridge, MA. (1999).
5. Dellarocas, C, Contractual Agent Societies: Negotiated shared context and social control in open multi-agent systems, Workshop on Norms and Institutions in Multi-Agent Systems, 4th International Conference on Multi-Agent Systems (Agents-2000), Barcelona, Spain, June (2000).
6. Galstyan, A., Kolar, S., and Lerman, K., Resource allocation games with changing resource capacities, Proceedings of the second international joint conference on Autonomous agents and multiagent systems AAMAS-03, Melbourne, Australia, (2003) 145–152.
7. Kumar, A., van der Aalst, WMP, and H.M.W. Verbeek, Dynamic Work Distribution in Workflow Management Systems: How to Balance Quality and Performance?, Journal of Management Information Systems, 18(3):157–193, (2002).
8. Jain, A.K., Aparicio, M. and Singh, M.P., Agents for Process Coherence in Virtual Enterprises. in Communications of the ACM, Volume 42, No 3, March (1999), pp62–69.
9. Wooldridge, M., Multiagent Systems, Wiley, (2002).

Learning Plans with Patterns of Actions in Bounded-Rational Agents

Budhitama Subagdja and Liz Sonenberg

Department of Information Systems
University of Melbourne
subagdja@pgrad.dis.unimelb.edu.au, l.sonenberg@unimelb.edu.au

Abstract. This paper presents a model of a learning mechanism for situated agents. The learning is described explicitly in terms of plans and conducted as intentional actions within the BDI (Beliefs, Desires, Intentions) agent model. Actions of learning direct the task-level performance towards improvements or some learning goals. The agent is capable of modifying its own plans through a set of actions on the run. The use of domain independent patterns of actions is introduced as a strategy for constraining the search for the appropriate structure of plans. The model is demonstrated to represent Q-learning algorithm, however different variation of pattern can enhance the learning.

1 Introduction

The term machine learning usually refers to the ability of a computer program to improve its own performance with experiences based on a performance measure [7]. One can view the learning as an extensive search for the appropriate knowledge that can be used to fulfill the performance measure criterion. The searching can be done in many ways such as manipulating connections, statistical measures, inductions, or analytical inferences. The result of the search process represents some regularities of the sampled experiences. Consequently, a situated agent should learn by identifying regularity from its experience in interacting with the environment. The agent then would reconfigure its own knowledge accordingly to reach certain criteria.

Dealing with changing circumstances has been a major issue in building situated rational agents. The BDI (Belief-Desire-Intention) agent architecture is a model for building agents which addresses the bounded rationality issue: deciding to do something in the environment, where time is limited, knowledge is inadequate, and resources are scarce. The BDI agent model provides a high-level programming abstraction for developing goal-directed reactive agents [9]. It has been formalized [2, 10, 13] and successfully applied in many domains of application with several types of frameworks and development tools(e.g. PRS, JACK, JAM) [4–6].

On the other hand, learning is not much considered in the BDI model. Inspired by philosophical concepts of intentions and practical reasoning [1], the

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 30–36, 2005.

BDI agent operates by executing plans as sequences of actions while deliberating. Unlike traditional planning, plans are not built from scratch each time the agent faces problems but they are pre-specified prior to the agent's execution. This approach makes the agent reactive and responsive to the changes in the environment within the limits of the pre-specified plans. It avoids planning time which could make the agent lag behind as significant changes happen during the planning. The adaptation is considered mostly on the issue of how the agent deliberates when certain events happened in the environment, but plans are still fixed inside their own repository.

This paper introduces a model of learning in BDI agent which conforms with the principle of bounded rationality. Learning is regarded as pre-specified plans executed in goal-directed but reactive manner. Furthermore, it is suggested that patterns of actions in plans can be used as heuristics to enhance learning. The patterns are domain independent cues which can guide the process of acquiring knowledge about regularities in the agent's interaction with the environment. Examples of the model show that Q-learning, a commonly-used reinforcement learning algorithm, can be described as plans in a BDI agent. However, it is also revealed that a standard way of modeling situation in Q-learning is insufficient for dealing with a dynamic environment. A work around by applying presumed patterns of actions can be used to reduce the complexity of the learning. A preliminary experiment has been conducted which demonstrates those matters.

2 The Agent Model

The BDI agent architecture is based on an intuitive model of human practical reasoning which capture changing states of reasoning in terms of explicit notions of *beliefs*, *goals* or *desires*, and *intentions*, about what actions should be taken next. The *beliefs* can be regarded as a database of facts representing states about the world that the agent realizes. Depending on the situation, the content of *beliefs* can be false or subject to change. *Goals* or *desires* represent states that the agent wants to bring about. Based on *beliefs*, the agent selects the most achievable goal to be realized and creates *intention* as a form of commitment towards executing actions that leads to the goal. The actions taken are based on *plans* as specifications of actions sequence to achieve the goal. For each execution cycle, the agent's reasoning is based on the interactions between those attributes.

The BDI agent architecture can be implemented as an interpreter. The interpreter goes through cycles of observing the environment, deliberating intentions, and executing actions. Desires or goals are obtained from pre-defined specifications or from external sources (such as requests from the user or particular events in the environment). When certain events occur implying changes in beliefs, plans applicable to those beliefs and goals are triggered. Some of these applicable plans then are selected and put on the intention structure. Finally, the interpreter selects an action from the root of the intention structure and executes it which results in either performing a primitive action, changing beliefs, or establishing new goals.

A plan generally specifies how to fulfill a particular goal at a certain condition. It has some specifications of the invoking goal and conditions which can make it applicable. The plan also has a *body* which describes the sequence of actions or procedure to be executed. A plan is successfully executed when all its intended actions are executed successfully. A plan or an action failure is inherently handled in the BDI mechanism by applying another applicable plan or dropping the goal and assuming that the goal can not be achieved. A plan can have other types of attribute like implied *effects*, *failure* handler, *utility*, or *cost*.

In the PRS (Procedural-Reasoning-System) architecture [6], one implemented BDI agent architecture, the process of deliberation and selection can be handled by meta-level plans. Meta-level plans are represented in the same way as domain-level plans. However, they operate on the agent's internal conditions like beliefs, goals, intentions, and plans themselves. A meta-level plan can represent complex decision-making and planning procedures. With the same kind of representation as any other plan, the execution of a meta-level plan can be interleaved with others as they all depend on the intention structure. We adopt this meta-level perspective for learning.

3 The Learning Model

This paper suggests that learning in BDI agent can be driven by meta-level plans. The idea is that meta-level plans are not used only for controlling deliberation and means-end reasoning but also to monitor executions of other plans and to adjust them when necessary. In contrast with other existing approaches of making BDI agents learn [3, 8], this approach put learning as deliberated reactive reasoning. Changes made from the learning would include updates on beliefs and plans.

The learning in the BDI agent is assumed to be conducted by monitoring experiences for certain patterns or regularities through events and the intention structure, and then accordingly manipulating plans to get a better performance. The process of monitoring and plan manipulation are driven by meta-level plans. The meta-level plans are not only invoked when the time for deliberation has come, but they can also be pursued when the domain-level plans are still running.

To enable such kind of learning with meta-level plans, several types of operators which support learning can be defined as follows:

1. **Monitoring operators.** This type of operator is used for monitoring events and intention properties at runtime. The operators must be able to capture events which indicate changes in the intention structure and their references to corresponding beliefs, goals, and plans.
2. **Plan construction and revision operators.** This type of operator is used for manipulating the content of plans in the plan repository. The operators can access all parts of a plan and manipulate them. They can be used also to construct a new plan from scratch.

3. **Deliberation operators.** This type of operator supports the deliberation process. There are operators that retrieve plans from the plan library based on some conditions of goals, beliefs, or intentions. There are also operators that select and bind a plan from given applicable plans based on certain criteria and put it on the intention structure.

Using those types of learning operator together with the expressive power of plan representation, any type of learning algorithm can be specified within a BDI agent. However, it is suggested that the process of learning goes through these steps:

1. **Generate a hypothesis.** In the execution of the meta-level plan, a hypothesis is made which assuming certain patterns of condition that will occurs in the environment or in the internal state of the agent.
2. **Test the hypothesis.** The hypothesis created is tested for its occurrences. The process of testing could involve steps waiting for certain events, or querying beliefs.
3. **Change or create plans.** After the hypothesis is confirmed, a plan can be constructed or revised. The construction or revision includes updating plan attributes (goals, precondition, context, effects, failures, etc.) or the procedural description in its body.

The performance measure to be improved can be explicitly specified as learning goals. These goals will invoke learning plans when certain conditions occur. In this way, learning is also treated as any other plan which can be deliberated and interleaved with each other. For example, Q-learning, as one of the most common approaches to reinforcement learning[12], can be specified to have an objective of maximizing rewards. The result of the learning process will be the optimal policy π^* for selecting action a from a certain state s:

$$\pi^*(s) = \arg\max_a Q^*(s, a) \tag{1}$$

To obtain the Q-function for the optimal policy, an update rule can be defined as follows:

$$Q(s_i, a_i) \leftarrow r_i + \gamma \max_{a \in A} Q(s_{i+1}, a) \tag{2}$$

in which r_i is the immediate reward from the environment after doing action a_i from state s_i, and γ is the discount factor that represents the relative importance of future rewards. The equation and rule of Q-learning shows that the learning will incrementally fills up a Q-table specifying optimal paths from any state to a direct rewarding state.

It is also possible to describe learning that exhibit the properties of Q-learning as a plan. The reward can be based on whether or not the plan execution has reached the goal. Successful execution will give a positive reward and a failure will produce the negative one. A performance measure can be obtained by setting up a reactive plan that always compare the sum of positive and negative rewards. The learning plan will be invoked when the measure reach a certain level. In this case, there must also be some pre-existing plans representing default actions.

The learning, firstly, must generate some hypotheses. A hypothesis frames some expectations about patterns that can occur on the agent activities. Patterns for the Q-learning describes the structure of plan bodies that will be generated from the learning process. Fig. 1 shows two patterns of Q-learning in UML activity diagram which describe the structure of plans that will be generated. In part (i) of Fig. 1, the structure of a plan that maps state s to the goal state x through the action a is shown. The part (ii) of Fig. 1 shows the structure of a plan that maps intermediate state s' to another state s through action a'. The plan will also recursively establishes x as its new subgoal. These patterns also imply that each plan generated resembles each entry on the Q-table.

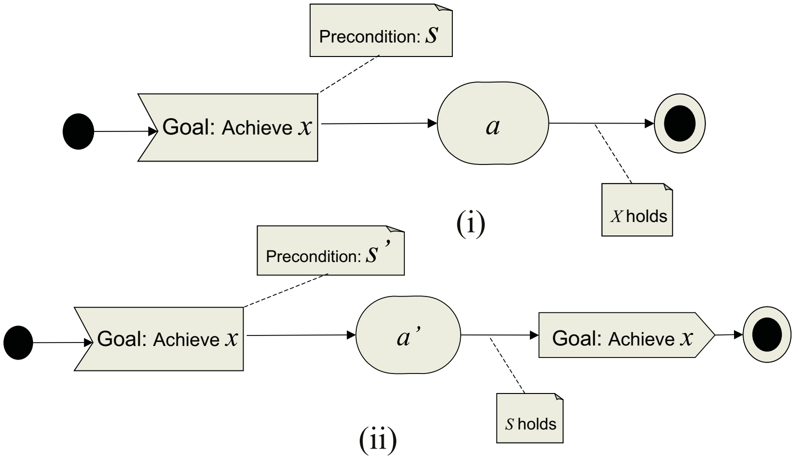


Fig. 1. Patterns for standard Q-learning

Secondly, the learning plan must recognize the hypothesis by monitoring events and changes in the environment and the intention structure. Finally, once a pattern is recognized a new plan can be created or an existing one can be modified depending on the procedural description on the corresponding meta-level plan. Fig. 2 shows a meta-level plan for creating plans based on the patterns for Q-learning plans. It must be admitted that some detail aspects of Q-learning (e.g. Q-values updates) are still not considered. However, it is assumed that those aspects can be handled by describing more complex patterns.

Although Q-learning can be implemented within the BDI agent, a problem still remains in finding the appropriate representation for states. An empirical investigation with a simple simulation discussed on the next section reveals that problem. In changing situations, attributes from the environment are still not sufficient to make the learning converge. The state must include all aspects of the environment combined with responses from environment after conducting certain actions.

To hinder from the combinatorial explosion in number of states, the plan representation in BDI agent can provide a way around. Instead of strictly adopt

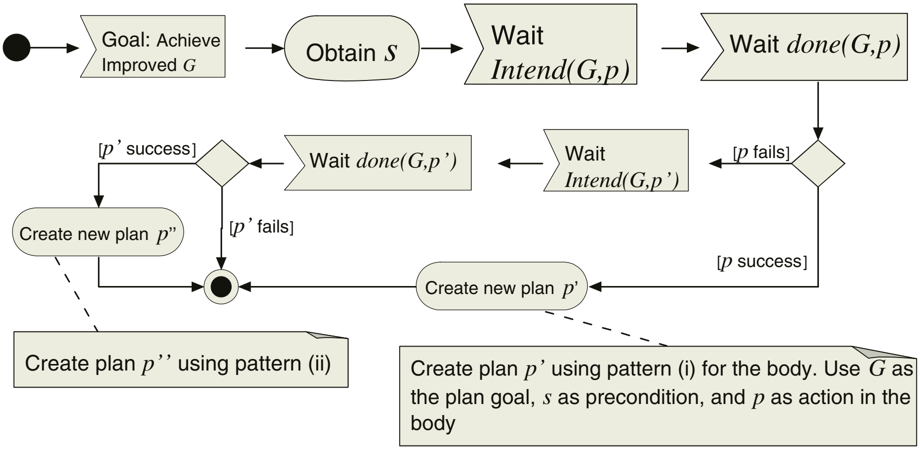

Goal: Achieve Improved G
Obtain S
Wait Intend(G,p)
Wait done(G,p)
[p fails]
[p success]
Wait Intend(G,p')
Wait done(G,p')
[p' success]
[p' fails]
Create new plan p''
Create new plan p'
Create plan p'' using pattern (ii)
Create plan p' using pattern (i) for the body. Use G as the plan goal, s as precondition, and p as action in the body

Fig. 2. The meta-level procedure for Q-learning

the pattern from some ad-hoc algorithms, some complex forms of pattern can be used as hypotheses. The pattern can consist of arbitrary sequences, branches, loops, or even recursions. Fig. 3 shows a pattern that can substitute the standard Q-learning pattern in part (i) of Fig. 1. Here, the pattern represents a short-cut in which the agent can assume that the goal can be achieved directly by applying two different consecutive actions. By this pattern, the agent is like presuming the behavior of the environment in the context of its interaction. This approach is suggested could reduce the complexity of the searching process in learning.

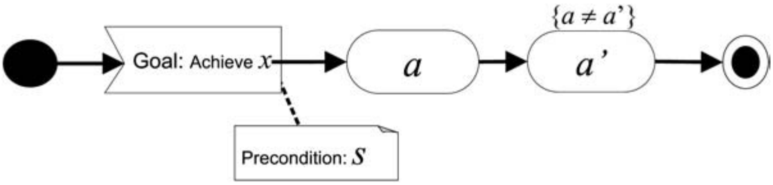


Fig. 3. A complex form of pattern for the hypothesis

4 Discussion

A simulation for testing and demonstrating the learning approach described above has been implemented [11]. An artificial agent is given a task of finding appropriate sequences of action that lead it to the goal. The learning task is non-trivial as the observable situation keeps changing. The experiment shows that the standard Q-learning cannot keep up with the changing paces and it settles on a condition similar to no learning at all. In contrast, learning with patterns like in Fig. 3, shows a good performance as the agent can find the right plans quickly. This result does not suggest that a reinforcement learning based algorithm can not learn a correct policy at all. There is always a work around by representing the environment in certain ways so that the agent can associate states and actions in more straightforward manner. However, designing the algorithm with its corresponding representation also means that the designer has already made an assumption about the model of relevant features of the environment which rather should be learnt by the agent. Moreover, patterns of hypothesis, which can also be regarded as learning heuristics, are considered as the main factor that influence the effectiveness of the learning.

References

1. Michael E. Bratman. *Intention, Plans and Practical Reason.* Harvard University Press, Cambridge, 1987.
2. Philip R. Cohen and Hector J. Levesque. Persistence, intention, and commitment. In Phillip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 33–69. MIT Press, Cambridge, 1990.
3. A. G. Hernandez, A. E. Segrouchini, and H. Soldano. BDI multiagent learning based on first-order induction of logical decision trees. In S. Oshuga N. Zhong, J. Liu and J. Bradshaw, editors, *Intelligent Agent Technology: Research and Development.* World Scientific, New Jersey, 2001.
4. Nick Howden, Ralph Ronnquist, Andrew Hodgson, and Andrew Lucas. Jack intelligent agents-summary of an agent infrastructure. In *Proceedings of the 5th International Conference on Autonomous Agents.* Montreal, 2001.
5. Marcus J. Huber. Jam: A BDI-theoretic mobile agent architecture. In *Proceedings of the Third International Conference on Autonomous Agents*, Seattle, 1999.
6. F. Ingrand, M. Georgeff, and A. Rao. An architecture for real-time reasoning and system control. In *IEEE Expert*, volume 7(6), pages 34–44. 1992.
7. Tom Mitchell. *Machine Learning.* McGraw-Hill, Cambridge, 1997.
8. Cindy Olivia, Chee-Fon Chang, Carlos F. Enguix, and Aditya K. Ghose. Case-based bdi agents: an effective approach for intelligent search on the world wide web. In *Proceedings of the AAAI-99, Spring Symposium on Intelligent Agents in Cyberspace Stanford University.* USA, 1999.
9. Anand S. Rao and Michael P. Georgeff. BDI agents: From theory to practice. In *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95).* San Francisco, 1995.
10. A.S. Rao and M. P. Georgeff. Modeling rational agents within a bdi-architecture. In R. Fikes and E. Sandewall, editors, *Proceedings of Knowledge Representation and Reasoning*, pages 473–484, San Mateo, 1991. Morgan Kaufmann.
11. Budhitama Subagdja and Liz Sonenberg. Reactive planning through reflective hypotheses: Towards learning in BDI agents. Submitted for AAMAS, 2005.
12. Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, 1998.
13. Michael Wooldridge. *Reasoning about Rational Agents.* MIT Press, Cambridge, 2000.

Roles of Agents in Data-Intensive Web Sites

Ali Ben Ammar[1], Abdelaziz Abdellatif[2], and Henda Ben Ghezala[3]

[1] Institut Supérieur de Documentation, Manouba 2010, Tunisia
ali.benammar@isd.rnu.tn
[2] Faculté des Sciences de Tunis, Tunis 2092, Tunisia
abdelaziz.abdellatif@fst.rnu.tn
[3] Ecole Nationale des Sciences de l'Informatique, Manouba 2010, Tunisia
henda.bg@cck.rnu.tn

Abstract. The Data-Intensive Web Sites provide access to a large number of Web pages whose content is dynamically extracted from structured databases. They may be used for shopping or paying in e-commerce, for reading news in a newspaper site or to consult digital library. In this context, users often need rich content and fresh data. Several techniques have been developed to meet the demand for faster and more efficient access to the DIWS. Among them a main role is acquired by the replication, the caching, the materialization, and the refreshing of data. Software agents have proved to be a good tool that may give a high performance results on the Web. In this paper, we address their cases of use in the DIWS. We discuss cases in which agents can be used to improve the data management performance. The aim is to specify tasks that may profit from the increase development in agent technologies.

1 Introduction

The Data-Intensive Web Sites (DIWS) provide access to a large number of Web pages whose content is dynamically extracted from structured databases. They serve to integrate and summarize Web services that may be distributed and heterogeneous. They may be used for shopping or paying in e-commerce, for reading news in a newspaper site or to consult digital library. Their source databases are generally distributed, heterogeneous, and with dynamic content. The user queries are, sometimes, personalized that is they are tailored to the style and the needs of each individual. They often demand rich content and fresh data. In this context, data management will be highly complex. It consists in integrating, updating and rapidly accessing data.

Several techniques have been developed to meet the demand for faster and more efficient access to the DIWS. Among them a main role is acquired by the replication, the caching, the materialization, and the refreshing of data. Software agents are demonstrated to be a good tool that may give high performance results in such environment [40,48,51,16]. An agent is a software entity, situated in an environment, where it acts autonomously and flexibly to reach some objectives [39]. A multi-agent system is a distributed system composed of a set of collaborative agents. It is used to perform distributed tasks.

In this paper, we address the use of agents in the DIWS. We discus the cases in which agents can be used to improve the data management performance. The aim is to specify tasks that may profit from the increase development in agent technologies. Our discussion is based on the study of some works involving agents to resolve data management problems.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 37–45, 2005.

The paper is organized as follows. Section 2 presents some related works. Section 3 presents the concept of DIWS. Section 4 describes in more detail the concept of agent. In section 5 we discuss the use of agents in some tasks of data management in DIWS. Section 6 concludes.

2 Related Works

Recently, there has been a lot of interest in the use of software agents in several domains. The main applications in which intelligent agents can be involved are identified in [37]. A framework to integrate agents into the use of the World Wide Web is designed and implemented in [16]. In this framework, the agents filter information, initiate communication, monitor events, and perform tasks. The aim is to improve the usability and usefulness of the World Wide Web. [48] surveys several agent-mediated e-commerce systems and [41] presents an example of such systems that automate the hotel reservation in tourism domain. In this paper we will limit our concern to the DIWS domain. By studying these related works and others we will identify the data management tasks which need software agents to be optimized.

3 Data-Intensive Web Sites (DIWS)

Data-Intensive Web Sites (DIWS) provide access to a large number of Web pages whose content is dynamically extracted from structured databases [28]. Today, they become necessary for allowing some e-commerce tasks or to access dynamic information. Their architecture includes a database management system (DBMS) layer, a site server layer and the client. Thus, a new kind of pages, dynamically generated, and a new architecture were born. We have no more the traditional couple of a Web client and a Web server, but a third part is added, the application program, running on the Web server to receive the request from the client, retrieve the relevant data from the database and then pack them into HTML or XML format. Newspaper sites and shopping ones are examples of such architecture. Several tools and approaches for developing such systems are presented in [43]. For each kind of application, a set of web development tools are specified. The performance problem of DIWS lies in addressing the latency reduction of page produced by the site and the quality of data presented to the clients. Firstly, because returning Web page may require costly interaction with the database system. So, the net effect of this situation is network congestion, high client perceived latency, Web server overload and slow response times for Web severs. Secondly, because the quality of data is of crucial importance, especially for applications that must always serve fresh data (e.g. providers of stock prices, sports scores).

Recently, much research has been devoted to improving Web performance by reducing latency and bandwidth consumption, and increasing server scalability and availability. Proposed solutions include predictive prefetching, caching and materialization of Web objects, and architecting network and Web servers for scalability and availability. These solutions are beneficial but need to be yet improved to accommodate the continuously growing number of web users and services. In section 5 we will discuss the possibility of using agents to improve these solutions.

4 Agents

According to [51], an agent is computational entity which:

- Acts on behalf of others entities in an autonomous fashion;
- Performs its actions with some level of proactivity and/or reactiveness ;
- Exhibits some level of the key attributes of learning, co-operation and mobility.

Software agents are, today, more important because:

- More and more every-day tasks are computer-based;
- The world is in a midst of an information revolution, resulting in vast amount of dynamic and unstructured information;
- Increasingly, more users are untrained;
- And therefore users require agents to assist them in order to understand the technically complex world we are in the process of creating.

According to [51], a mobile agent is a software entity which exists in a software environment. It inherits some of the characteristics of an agent. A mobile agent must contain all of the following models: an agent model, a life-cycle model, a computational model, a security model, a communication model and finally a navigation model.

According to [51], a multi-agent system is able:

- To solve problems that are too large for a centralised single agent to deal with due to resource limitations or the sheer risk of having one centralised system;
- To allow for the interconnecting and interoperation of multiple existing legacy system, e.g. expert systems, decision support systems;
- To provide solutions which draw from distributed information sources;
- To provide solutions where the expertise is distributed, e.g., in health care provisioning;
- To enhance speed (if communication is kept minimal), reliability (capability to recover from the failure of individual components with graceful degradation performance), extensibility (capability to alter the number of processors applied to a problem), the ability to tolerate uncertain data and knowledge;
- To offer conceptual clarity and simplicity of design.

In section 5 we will give examples that illustrate these abilities. Then, we will deduce where and how agents can be used in DIWS to make profit from their abilities.

5 Improving DIWS Performance by Using Agents

As we have seen above, several solutions have been developed to improve web performance. In major cases, these solutions are still valid for DIWS. We may classify them into three groups: data integration solutions, data update solutions, and data access optimization solutions. In the rest of this section we will see how these solutions can be improved by using agents.

5.1 Data Integration

To construct a client web page in DIWS environment, data should be extracted from different sources and then integrated. The integration needs metadata that describe the data semantic and the mapping approach from database to Web page. Several techniques of semantic Web like ontology [60] have been used to perform the integration. So, the integration needs three main tasks: searching metadata, constructing and updating the mapping approach, and composing the Web page to the client.

Software agents have been used in several domains needing integration [13,15,33,38,49,51]. The main agent types that we see more relevant for the integration are:

- Wrapper agents that convert the source information and react to source changes ;
- Integrator agents that manage global data view, transform and subdivide queries, integrate and formulate responses.

Since source data and user queries are high dynamic in a DIWS, these two agent types may be more adequate to optimize the integration process. For complex queries asking replicated and distributed data, integration agents will greatly decrease their response time.

5.2 Data Update

Data update may concern the modification of source data, the refreshment of data copies, or the refreshment of metadata in an integrator. This subsection deals with only the refreshment of data copies since in DIWS environments the data sources are, in general, managed by their owners and that the metadata refreshment is evoked here above. Data copies mean the data extracted from a source to be integrated in a Web page which is materialized or cached in a Web proxy or in a Web server. There are many works addressing Web caching data [1,11,14,25,27,28,45,46] and Webview materialization [2,4,5,7,9,17]. A key requirement for DIWS that provide dynamic data is to keep data copies up-to-date that is fresh and consistent with their original sources. The freshness of data [31] depends, in general, on the client tolerance and on data access frequency.

On the Web, there are many techniques to refresh derived data [3,6,24,30,35,42] but there is less use of agents. This may be because there are database tools, like triggers, that can perform such task. In [35], there is an attempt to use agents for capturing source updates. Agents are used in [22,57] to capture user needs and preferences which may lead to deducing user tolerance. i.e. accepting a data that is not refreshed along a period of time t, means that the user is satisfy if the data age is less than t. So, in DIWS, agents may be used for two updating tasks: capturing source changes and specifying data update frequencies based on user tolerance and access frequencies.

5.3 Data Access Optimization

A family of optimization techniques is developed to reduce the query response time in DIWS environment. It includes distribution, data caching and data materialization.

5.3.1 Distribution

With the increase in traffic on the web, popular Web sites get a large number of requests. Servers at these sites are sometimes unable to handle the large number of requests and clients to such sites experience long delays. One approach to overcome this problem is the distribution or replication of content over multiple servers. This approach is called Content Distribution Networks (CDN). It allows for client requests to be distributed over multiple servers. Several techniques have been suggested to distribute or replicate content over Web servers [18,29,34,36,52,53,54,61,64], and to direct client requests to multiple severs [12,19,26,32,55,56].

Based on their intelligence and learning capability, software agents can optimize query response time essentially in distributed environment. They are used to collect information on the Web [10,23, 33,58,63]. In the DIWS context, query processing may be distributed over several mobile or source localised agents. Agents may also be used to dynamically searching the optimal processing plan for query in a given situation (server overload, data distribution,…).

5.3.2 Caching

DIWS satisfy, in most case, dynamic requests. The overhead for satisfying this kind of requests may be orders of magnitude more than the overhead for satisfying static requests. Dynamic requests often involve extensive back-end processing and invoke several database accesses. In order to reduce the overhead of generating dynamic data, it is often feasible to generate the data corresponding to a dynamic page once, store the page in a cache, and to serve subsequent requests to the page form cache instead of invoking the server program again. However there are types of dynamic data that cannot be pre-computed and serviced from the cache. For example a personalized Web page that contains content specific to a client, such as the client's name, should not be cached.

The issues pertaining to the cache management are cache consistency and cache replacement. The purpose of a cache consistency mechanism is to ensure that cached data are essentially updated to reflect the changes to the original data. While, the purpose of a cache replacement mechanism is to decide which data will enter the cache when a client requests them and which data already in cache will be purged out in order to make space for the incoming data when the available space is not sufficient. The first issue is a data updating problem that is evoked in the subsection 5.2. We will now address the role of agents in the cache replacement problem and query processing.

Several replacement algorithms have been developed in literature [1,8,11,14, 27,44]. They try to keep in cache the most valuable data. The value of datum is usually a function of several parameters, say access frequency, size, retrieval cost, frequency of update etc…. In [11], the authors propose to use fragments to allow partial caching of dynamic pages. Common information that needs to be included on multiple Web pages can be created as a fragment. In order to change the information on all pages, only the fragments need to be changed. In this context agents may be used to search common fragments and then to identify the appropriate ones that should be cached.

Intelligent agents may be used also to prefetching Web pages that will be probably highly accessed in the next period. That is they will prevent the cache content before receiving queries. Agents may be used also to transform some cached data in order to satisfy incoming queries that don't have cached solutions. For this reason, agents should analyze the queries and profit from their experience to provide good responses.

Web data may be cached in several nodes of the network (DBMS, Web server, Proxy, …). In this case, agents may be distributed over the different nodes to manage caches. Their role will be to negotiate the relevant data to be cached, when to place data, how to compose the query responses from the distributed fragments.

5.3.3 Materialization

Similarly to traditional database views, the term Webviews is used on the web to mean Web pages that are automatically constructed from base data using a program or a DBMS query. The materialization approach consists in computing Webviews and storing them. Having a Webview materialized can potentially give significantly lower query response times, provided that the update workload is not heavy. Even if the Webview computation is not very expensive, by keeping it materialized we eliminate the latency of going to the DBMS every time which could lead to DBMS overloading.

According to [5], Webview materialization is different from Web caching: Webview materialization aims at eliminating the processing time needed for repeated generation, whereas Web caching strives to eliminate unnecessary data transmissions across the network.

The Webview materialization approach is similar to that of view materialization in a data warehouse [20,21,47]. The main issues of the Webview materialization approach are: how to select dynamically the appropriate Webviews to be materialized, how to refresh materialized Webviews and how to distribute the storage of Webviews over several servers.

Here, agents may be used in the first task to search the needed information and parameters like the access frequency, the update frequency, the estimated size of Webviews; to decide which Webviews to materialize in a given situation (reserved space, overload constraints,…). The role of agents in the two other tasks will be as it is described in sections 5.3.1 and 5.2. In the query processing context, agents can reformulate query to be satisfied from the materialized Webviews or redirect query to the appropriate server having the responsive Webviews.

6 Conclusion

Today, software agents are frequently used on the Web to optimize several data management tasks. In this paper, we have addressed their role in DIWS. After describing the concepts of DIWS and agents, we have identified the main tasks of DIWS, in which agents can be involved. From the study of some applications of agents on the Web, we have concluded that, in a DIWS environment, software agents can enhance the performance of other techniques developed to perform three main functions: data integration, data update and data access optimization. In the future work we will, develop a multi-agent system that dynamically selects the appropriate Webviews to be materialized.

References

1. Iyengar, D. Rosu. Architecting Web sites for high performance. Scientific Programming 10(1): 75-89. 2002
2. Labrinidis, N. Roussopoulos. Adaptive WebView Materialization. WebDB'01: 85-90. 2001
3. Labrinidis, N. Roussopoulos. Balancing Performance and Data Freshness in Web Database Servers. VLDB'03: 393-404. 2003

4. Labrinidis, N. Roussopoulos. Online View Selection for the Web. 2002. http://citeseer.ist.psu.edu/cache/papers/cs/25933/http:zSzzSzwww.cs.umd.eduzSzLibraryzSzTRszSzCS-TR-4343zSzCS-TR-4343.pdf/labrinidis02online.pdf
5. Labrinidis, N. Roussopoulos. On the Materialization of WebViews. In ACM SIGMOD Workshop on the Web and Databases (WebDB '99): 79-84. 1999.
6. Labrinidis, N. Roussopoulos. Update Propagation Strategies for Improving the Quality of Data on the Web. VLDB'01: 391-400. 2001
7. Labrinidis, N. Roussopoulos. WebView Materialization. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. pp.367-378. May 15-18, 2000, Dallas, Texas, United States
8. Labrinidis. Web-Aware Database Servers – I. 2002 http://www.cs.pitt.edu/~labrinid/courses/cs2001/webdb-12nov2002.pdf
9. Christos, K. Agisilaos. Efficient Materialization of Dynamic Web Data to Improve Web Performance. 15th International Conference on Computer Communication (ICCC 2002) Mumbai.India, August 11-14 2002
10. Edgar, G. Susan. Intelligent Information Agents for the World Wide Web. Technical Report ITTC-FY97-TR-11100-1, Information and Telecommunication Technology Center, The University of Kansas, Lawrence, KS, May 1997.
11. Mohan. Caching Technologies for Web Applications. In Proceedings of the 2001 VLDB Conference, Roma, Italy, September 2001.
12. A. Johnson, G. C. Shoja. Request Routing Architecture in Content delivery Networks. Proceedings of 2003 International Conference on Internet Computing, June 23-26, 2003, Las Vegas, USA.
13. D. Gilbert, M. Aparicio, B. Atkinson, S. Brady, J. Ciccarino, B. Grosof, O'Connor, P., Osisek, D., Pritko, S., Spagna, R. & Wilson L. The role of intelligent agents in the information infrastructure. IBM Report. 1995.
14. D. Katsaros, Y. Manolopoulos. Cache management for Web-powered databases. In Web-Powered Databases, pp 201-242. IDEA Group Publishing. 2002.
15. Scilla, M.N. Huhns. Making Agents Secure on the Semantic Web. IEEE Internet Computing, pp 76-93, Nov/Dec 2002.
16. Fischer, C.G. Thomas. Using Agents to improve the Usability and Usefulness of the World- Wide Web. Proceedings UM-96, Fifth International Conference on User Modeling, Hawaii, User Modeling, Inc. 1996, 5-12.
17. Mecca, A. O. Mendelzon, P. Merialdo. Efficient Queries over Web Views. EDBT '98: 72-8. 1998
18. G. Pierre, M. van Steen, A. S. Tannenbaum. Dynamically Selecting Optimal Distribution Strategies for Web Documents. IEEE Transactions on Computers, v.51 n.6, pp 637-651, June 2002
19. Kabir, E. G. Manning, G. C. Shoja. Request-Routing Trends and Techniques in Content Distribution Network. Proc ICCIT 02, Dhaka, Bangladesh, pp 315-320. December 2002. ISBN 984-32-0450-6.
20. H. Gupta and I.S. Mumick. Selection of Views to Materialize Under a Maintenance-Time Constraint. ICDT. 1999.
21. H. Gupta. Selection of Views to Materialize in a Data Warehouse. ICDT. 1997.
22. H. Liebermann. Letizia: An agent that assists Web browsing. In *Proc. Intl. Conf. on AI*, Montréal, Canada, August 1995.
23. H. Lieberman, N. van Dyke, A. Vivacqua. Let's browse: a collaborative Web browsing agent. In Proc. Intl. Conf. on Intelligent User Interfaces, January 1999.
24. Ari, E. L. Miller. Caching support for push-pull data dissemination using data snooping routers. In Proceedings of the 10th International Conference on Parallel and Distributed Systems (ICPADS’04). 2004

25. Wang. A Survey of Web Caching Schemes for the Internet. ACM Computer Communication Review 29(5). 1999.
26. Watts, S. Taylor. A Practical Approach to Dynamic Load Balancing. IEEE Trans. on Parallel and Distributed Systems, Vol. 9, NO. 3, March 1998, pp 235-248.
27. Amiri, S. Park, R. Tewari, S. Padmanabhan. DBProxy: A Dynamic Data Cache for Web Applications. ICDE Conference 2003. pp 821–831.
28. Yagoub, D. Florescu, V. Issarny, and P. Valduriez. Caching strategies for data intensive web sites. In Proceedings of the VLDB 2000 Conference, pp 188-199. 2000.
29. Lei Gao, Michael Dahlin, Amol Nayate, Jiandan Zheng, Arun Iyengar. Improving Availability and Performance with Application-Specific Data Replication. IEEE Trans. Knowl. Data Eng. 17(1): 106-120. 2005
30. Bhide, K.Ramamritham, And P. Shenoy. Efficiently Maintaining Stock Portfoilis Up-To-Date On The Web. IEEE Reasearch Issues In Data Engineering (RIDE'02) Workshop, March 2002.
31. Bouzeghoub, V. Peralta. A Framework for Analysis of Data Freshness. IQIS 2004. pp 59-67.
32. Conti, E. Gregori, F. Panzieri. Load distribution among replicated web servers: A QOS based approach. in Second Workshop on Internet Server Performance, May 1999.
33. M. Côté, N. Troudi. NetSA: Une Architecture Multiagent pour la Recherche sur Internet. Expertise Informatique, vol. 3(3). 1998.
34. M. Karlsson, C. Karamanolis. Choosing Replica Placement Heuristics for Wide-Area Systems. Proceedings of the 24th International Conference on Distributed Computing Systems (ICDCS'04), pp 350-359, March 24-26, 2004
35. Ashish, D. Kulkarni and Y. Wang. Source Update Capture in Information Agents. Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03), Acapulco, Mexico. August 9-10, 2003.
36. N. A. John. A Study of Replicated and Distributed Web Content. Thesis, Departement of Computer Science, Worcester Polytechnic Institute, August 2002.
37. http://www.wpi.edu/Pubs/ETD/Available/etd-0810102-160719/unrestricted/john.pdf
38. N. Jennings, M. Wooldridge. Applications of intelligent agents. Chapter 1 in Agent Technology: Foundations, Applications, and Markets. Springer, 1998.
39. N. Gibbins, S. Harris, N. Shadbolt. Agent-based Semantic Web Services. WWW2003, May, 2003.
40. N.R. Jennings, M. Wooldrige, K. Sycara. A Roadmap of agent research and development. Int Journal of Autonomous Agents and Multi-Agent Systems, 1(1):7-38,1998.
41. N.R.Jennings. An Agent-based approach for building complex software systems. Communications of the ACM, Vol.44, No.4. April 2001.
42. Dikenelli, N.Y. Topaloglu, C. Erdur, O. Ünalir. An Agent-Mediated E-Commerce Framework For Tourism Domain.
http://www.srdc.metu.edu.tr/webpage/projects/hermesproject/documents/agent-tourism.pdf
43. Deolasee, A. Katkar, A. Panchbudhe, K. Ramamritham, and P. Shenoy. Adaptive push-pull: Disseminating dynamic web data. In Proceedings of the 10th International World Wide Web Conference, pages 265-274, Hong Kong, China, May 2001.
44. P. Fraternali. Tools and Approaches for Developing Data-Intensive Web Applications: A Survey. In ACM Computing Surveys Volume 31:3, pp 227-263. 1999.
45. Luo, J. F. Naughton, R. Krishnamurthy, P. Cao, Y. Li. Active Query Caching for Database Web Servers. WebDB 2000. pp 29-34. 2000.
46. Q. Luo, J. Naughton. Form-based proxy caching for database-backed web sites. In Proceedings of the 2001 VLDB Conference, pages 191-200, September 2001.
47. Q. Luo, J. Naughton, R. Krishnamurthy, P. Cao and Y. Li. Active Query Caching for Database Web Servers. WebDB, 2000.

48. Chirkova, A. Halevy, D. Suciu. Formal Perspective on the View Selection Problem. In Proc. VLDB, Roma, Italy, pp 59-68. 2001.
49. R.Guttman, A.Moukas, P.Maes. Agent-mediated electronic commerce: A survey. Knowledge Engineering Review, 13(2):147-159, 1998.
50. R. J. Bayardo, W. Bohrer, R. Brice, A. Cichocki, J. Fowler, A. Helal, V. Kashyap, T. Ksiezyk, G. Martin, M. Nodine, M. Rashid, M. Rusinkiewicz, R. Shea, C. Unnikrishnan, A. Unruh, D. Woelk. Infosleuth: agent-based semantic integration of information in open and dynamic environments. In Huhns M. N. et Singh M. P., éditeurs, Reading in Agents, pages 205–216, SF, CA, 1998. Morgan Kaufmann.
51. Bergamaschi, G. Cabri, F. Guerra, L. Leonardi, M. Vincini, F. Zambonelli. Mobile Agents for Information Integration. 2001.
http://dbgroup.unimo.it/Miks/publications/papers/cia2001.pdf
52. S. Green, L. Hurst, B. Nangle, P. Cunningham, F. Somers, R. Evans. Software Agents: A Review. Intelligent Agents Group (IAG) report TCD-CS-1997-06, Trinity College Dublin, May 1997.
53. S. Sivasubramanian, M. Szymaniak, G. Pierre, M. van Steen. Web Replica Hosting Systems Design. Internal Report IR-CS-001, June 14, 2004.
http://www.cs.vu.nl/pub/papers/globe/IR-CS-001.03.pdf
54. Rotaru, H. Nägeli. Dynamic load balancing by diffusion in heterogeneous systems. Journal of Parallel and Distributed Computing, v.64 n.4, p.481-497, April 2004
55. Cardellini, M. Colajanni, P. S. Yu. Dynamic Load Balancing on Web-Server Systems. IEEE Internet Computing, v.3 n.3, p.28-39, May 1999
56. V.Cardellini, M. Colajanni, P. S. Yu. Request Redirection Algorithms for Distributed Web Systems. IEEE Transactions on Parallel and Distributed Systems, v.14 n.4, p.355-368, April 2003
57. Ghini, F. Panzieri, M. Roccetti. Client-centered Load Distribution: A Mechanism for Constructing Responsive Web Services. Proceedings of the 34th Hawaii International Conference on System Sciences – 2001.
58. V. Robles, E. Menasalvas, S. Millán, M. Pérez, E. Hochsztain, O.Marbán, J. Peña, A. Tasistro. Beyond user clicks: an algorithm and an agent-based architecture to discover user behaviour. Proceedings of the ECML/PKDD 22-26 September 2003. Cavtat. Dubrovnik, Croatia.
59. Jirapanthong, T. T. Sunetnanta. An XML-Based Multi-Agents Model for Information Retrieval on WWW. In Proceedings of the 4th National Computer Science and Engineering Conference (NCSEC2000), Queen Sirikit National Convention Center, Bangkok, Thailand, November 16-17 .2000.
60. W. Li, W. Hsiung, D. V. Kalashnikov, R. Sion, O. Po, D. Agrawal, K. S. Candan. Issues and Evaluations of Caching Solutions for Web Application Acceleration. In *Proceedings of the 28th Very Large Data Bases Conference.* Hong Kong, China, August 2002.
61. Zhu, S. Gauch, L. Gerhard, N. Kral, A. Pretschner. Ontology-Based Web Site Mapping for Information Exploration. Proc. 8th Intl. Conf. on Information and Knowledge Management (CIKM '99), pp. 188-194, Kansas City, MO. November 1999.
62. Chen, L. Qiu, W. Chen, L. Nguyen, R. H. Katz. Efficient and adaptive web replication using content clustering. IEEE J. Sel. Areas Commun. 21, 6 (Aug.), 979--994. 2003
63. Y. Kotidis, N. Roussopoulos. DynaMat: a dynamic view management system for data warehouses. ACM SIGMOD Record, v.28 n.2, p.371-382, June 1999.
64. Y. Xirouhakis, G.Votsis, K.Karpouzis, S.Kollias. Efficient Browsing in Multimedia Databases using Intelligent Agents and Content-Based Retrieval Schemes. IEEE International Workshop on Multimedia Signal Processing (MMSP'98), Los Angeles, California, USA.1998.
65. Y. Zhang, B. Krishnamurthy, C. Wills. On the Use and Performance of Content Distribution Networks .In ACM SIGCOMM Internet Measurement Workshop. 2001

Probabilistic Reasoning Techniques for the Tactical Military Domain

Catherine Howard[1] and Markus Stumptner[2]

[1] Electronic Warfare and Radar Division
Defence Science and Technology Organisation
PO Box 1500, Edinburgh, South Australia, 5111
catherine.howard@dsto.defence.gov.au
[2] Advanced Computing Research Centre, University of South Australia
Adelaide, South Australia, 5095
mst@cs.unisa.edu.au

Abstract. The use of probabilistic reasoning is a key capability in information fusion systems for a variety of domains such as military situation assessment. In this paper, we discuss two key approaches to probabilistic reasoning in military situation assessment: Probabilistic Relational Models and Object Oriented Probabilistic Relational Models. We compare the modeling and inferencing capabilities of these two languages and compare these capabilities against the requirements of the military situation assessment domain.

1 Introduction

Decision making in time-critical, high stress, information overloaded environments, such as the tactical military domain, is a complex research problem that can benefit from the application of information fusion techniques. Information fusion is the process of acquiring, aligning, correlating, associating and combining relevant information from various sources into one or more representational formats appropriate for interpreting the information. We have adopted the Lambert revision [1] (λJDL) of the widely accepted Joint Directors of Laboratories model [2] which provides a functional model of the information fusion process. λJDL divides the information fusion into three sub-processes: object, situation and impact fusion.

Our research focuses the use of automated reasoning techniques to produce situation assessments from signal data in the military domain to support tactical decision-making. Signal data provides only a partial picture of the battle space. It may be incomplete, incorrect, contradictory or uncertain and may have various degrees of latency. It may also be affected by the environment or by enemy deception or confusion, which creates false or misleading data. Within the λJDL model, situation assessments are defined as persistent representations of the relationships of interest between objects of interest in the area of interest in the battlespace [1]. Such relationships of interest can include physical, temporal, spatial, organizational, perceptual and functional relationships. The relationships meaningful to a user will be highly dependent on the domain and the user's intentions.

In order to derive Situation Assessments from signal data, we need to model the battlespace to reason about the location, status and the relationships which exist between military units in the battlespace. Such a model requires the ability to [3]:

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 46–53, 2005.

R1 Represent objects and their attributes

R2 Represent relationships and their attributes

R3 Handle the uncertainty present in the domain in a computationally robust and mathematically sound manner.

Types of uncertainty present in the domain include:

- **R3.1 Existence uncertainty:** uncertainty about the existence of objects and relationships in the battlespace
- **R3.2 Attribute uncertainty:** uncertainty about the attributes of an object or relationship (e.g. object location)
- **R3.3 Structural uncertainty:** which can be broken down into:
  - **R3.3.1 Number uncertainty:** uncertainty over the number of objects in the battlespace
  - **R3.3.2 Reference uncertainty:** uncertainty over relationships between objects in the battlespace
  - **R3.3.3 Identity uncertainty:** uncertainty over the identity of an object

R4 Reason about groups of objects or relationships

R5 Fuse information at various levels of abstraction

R6 Perform temporal reasoning

Bayesian Networks are a popular technique which have been used in the military domain to reason about causal and perceptual relationships between objects in the battlespace (e.g. [4-10]). However, they have recently been shown to be inadequate for reasoning about large, complex domains [4, 11] because of their lack of flexibility and inability to take full advantage of domain structure or reuse. For example, the battlespace involves an unknown number of objects interacting with each other in a variety of ways which evolve over time. When a model of the battlespace is designed, we do not know the objects that will be present, their attributes or the relationships in which the objects will participate. However, BNs are rigid: they model the domain with a predefined set of random variables and a fixed topology which applies to all problem instances of the domain. Hence, they cannot represent uncertainty about the existence, number or configuration of objects in the battlespace. This limitation is because BN lack the concept of an object. And because they have no notion of objects, BNs cannot take full advantage of the structure of the domain or reuse. This lack of flexibility is of particular relevance to the situation assessment application area because the variables relevant to reasoning about a situation are highly dependent on the domain and the user intentions.

From this discussion, it is clear that representation and reasoning techniques are required which allow the random variables in the model, their state spaces and their probabilistic relationships to vary over time and from instance to instance. The techniques should also meet requirements (R1-R6) and also take advantage of the structure of the domain and reuse. We have developed Object Oriented Probabilistic Relational Models (OPRMs) [3], a new First Order Probabilistic Languages (FOPLs) to address some of these requirements. Section 2 of this paper provides a basic outline of Object Oriented Probabilistic Relational Models. Section 3 discussed the inferencing techniques for this language and compares the techniques to inferencing techniques used for Probabilistic Relational Models, a closely related language. Section 4 outlines future work while the conclusion are presented in Section 5.

2 Object Oriented Probabilistic Relational Models

OPRMs are a probabilistic frame based representation language that allow the problem to be modelled in a structured manner in terms of the objects and the relationships between them. OPRMs overcome the main limitations of BN in complex domains; they are flexible, they enable compact temporal representations, and they take advantage of the domain structure and facilitate reuse. The thesis behind this work is that OPRMs will provide a flexible and practical approach to reasoning in complex domains such as tactical miliary situation assessment. The OPRM language is still under development, but as it stands, it is able to address requirements R1, R2, R4, R5 and R3.1, 3.2, 3.3.1 and 3.3.2.

OPRMs specify a template for the probability distribution over a knowledge base. This template consists of two parts: a relational component and a probabilistic component. The relational component describes how the classes in the domain are related. The probabilistic component details the probabilistic dependencies between attributes in the domain.

Definition: The relational component of an OPRM consists of:

- A set of classes, $\mathbf{C} = \{C_1, C_2, \ldots, C_n\}$, and possibly a partial ordering over **C** which defines the class hierarchy.
- A set of descriptive attributes **A** for each class C in **C**. C_1.A is an attribute A of class C_1. Each descriptive attribute A has a domain type Dom[A]$\in$**C** and a range type Range[A]=Val[A] where Val[A] is a predefined finite enumerated set of values. The set of descriptive attributes of class C is denoted *A*(X).
- A set of reference attributes ρ for each class C in **C.** C_1.ρ is a reference attribute ρ of class C_1. Reference attributes represent functional relationships between instances in the knowledge base (i.e. they are attributes which reference other frame instances). Each reference attribute ρ has a domain type Dom[ρ]$\in$**C** and a range type Range[ρ]$\in$**C** for some class C in **C**. The set of reference attributes of class C is denoted *R*(X).
- A set of named instances, **I**, which represent instantiations of the classes. As multiple inheritance is not accommodated in this language, each instance is an instance of only one class.

The probabilistic component of OPRM consists of:

- A set of conditional probability models P(*A*|Pa[*A*]) for the descriptive attributes, where Pa[*A*] is the set of parents of *A*. These probability models may be attached to particular instances or inherited from classes.

The development of OPRMs followed the principles of Probabilistic Relational Models which were first developed by [12] and later refined by [13, 14]. Both PRMs and OPRMs integrate a frame-based representation with a probability model representing the uncertainty over the attributes of an instance. This probability model captures the probabilistic dependence of the instance on both its own attributes and the attributes of related instances. Both OPRMs and PRMs translate this probability model into the equivalent Object Oriented Bayesian Network (OOBN).

Definition (following [15]): An OOBN class is a BN fragment containing *output*, *input*, and *protected* (or *encapsulated*) nodes. The input and output variables form the *interface* of the class. The interface encapsulates the internal variables of the class, d-separating them from the rest of the network. All communication with other instances is formulated in terms of probability statements over the instance's interface.

There are currently two OOBN frameworks: the framework developed by Koller and Pfeffer (KPOOBNs) [16] and the framework developed by Bangsø (BOOBNs) [15, 17-19]. While PRMs utilize KPOOBN, OPRMs utilize BOOBNs. The main difference between the two OOBN frameworks is that BOOBNs introduce the use of *reference nodes* and *reference links* to overcome the problem that no node inside a class can have parents outside the class. A reference node is a special type of node which points to a node in another scope (called the *referenced* node). A reference node is bound to its referenced node by a reference link. BOOBNs define all input nodes to be reference nodes.

While these reference nodes create an additional computational cost, they provide several important benefits such as enabling BOOBNs to compactly represent dynamic situations and enabling BOOBNs to have a more intuitive definition of inheritance. KPOOBNs cannot handle 'non-connected input' and because of this, an interface type, t', of class C', $t'=(I(t'),O(t')$, is a subtype of an interface type $t = (I(t),O(t))$ of class C if $I(t') \subseteq I(t)$ and $O(t) \subseteq O(t')$ [13]. This definition means that the set of interface attributes for the subtype interface $t'$is smaller than the set of interface attributes for interface type t. However, the set of output attributes for the subtype interface $t'$is larger than the set of output attributes for the interface type t. Because BOOBNs include reference nodes for the input nodes which can be assigned default potentials, they has a more intuitive inheritance definition where class C' is a subclass of C if $I(t) \subseteq I(t')$ and $O(t) \subseteq O(t')$.

The use of reference nodes also provides important benefits for inferencing.

3 Inference

In the following discussion on inferencing techniques, a knowledge base is defined as consisting of a set of classes, a set of instances a set of inverse statements and a set of instance statements.

3.1 PRM Inferencing Techniques

In their SPOOK system, Pfeffer et al [12, 20] present two different techniques for reasoning with PRMs: the Structured Variable Elimination (SVE) for reasoning with classes and the Knowledge Based Model Construction (KBMC) technique for reasoning with instances.

To solve a query on class C, using the knowledge base, the SVE algorithm constructs a local BN for C consisting of a node for each attribute of C in addition to special output and projection nodes. When the algorithm comes across a reference attribute in C (of type C′), it eliminates this attribute by performing a recursive call. The call generates a temporary local flat BN for the class C′. It makes use of the fact that the interface of C′ encapsulates the internal variables of the class, d-separating them from the rest of the network. Using an efficient variable elimination order based on the structure of the temporary local BN, the algorithm uses standard variable elimination techniques to compute the distribution over the interface of C′. This distribution is then used in the BN representing class C to solve the user's query.

By taking this structured approach to the elimination of reference attributes, SVE takes advantage of the structure of the domain. Recursively solving for the distributions over class interfaces requires each recursive computation to only *temporarily* instantiate the equivalent BN for the class. This leads to space and arguably time savings due to the ordering of elimination. However, it is worth noting that the SVE algorithm requires the introduction of output and projection nodes into the class model as a direct consequence of the fact that in PRMs, the interfaces of classes must be dynamically determined (because is not known apriori how one class depends on attributes of another). As a result, more nodes are introduced into the network than in OPRMs where reference nodes are used to clearly define class interfaces.

SVE takes advantage of reuse by maintaining a persistent cache of results of queries executed on the classes (caching is of little use when dealing with instances). By maintaining such a cache, computations across a query and also across different queries on the same model can be reused. Computations can also be reused between different models in which the same class appears.

Pfeffer et al [20] reason that when dealing with instances, there may be situations where the instance interfaces no longer encapsulate the protected attributes from the rest of the network, which means that the SVE technique can not be used. For *all* instances, they employ a KBMC technique to construct one flat BN containing all attributes of all the instances using a backward chaining algorithm. Inference is then performed using this BN. The SVE algorithm and the KBMC algorithm are similar in that they both use the knowledge base to construct a local BN upon which inferencing is performed and they both employ backward chaining techniques to construct their models. However, the KBMC algorithm fails to take advantage of the structure of the domain (all structure is lost as soon as the model is translated into one flat BN) or reuse. Thus when dealing with a knowledgebase that contains a large proportion of classes compared to instances, Pfeffer et al's approach would have a computational advantage over traditional BN techniques because the SVE algorithm would perform the majority of the reasoning. However, in the situation where there are a large proportion of instances in the knowledgebase, while the modeling techniques would retain their representational benefits over BNs, the inferencing techniques would provide little benefit over traditional BN techniques as they fail to take advantage of the structure of the domain or reuse.

3.2 OPRM Inferencing Techniques

Inference in OPRMs is performed by dynamically constructing the 'equivalent' BOOBN for each class in the knowledgebase. The protected nodes in these equivalent BOOBN are encapsulated from the rest of the model via the class's interface and the inference algorithms take advantage of this fact. One of the advantages of the BOOBN framework is that the use of reference nodes means that the interfaces of each class are fully specified once the classes are created (i.e. at design time) and therefore do not need to be dynamically determined. As such, we are able to take advantage of Bangsø's techniques [15, 17] to translate the network directly into junction trees rather than into the equivalent flat BN and thence into a junction tree. This enables each class can be ‘precompiled’ into a junction tree once the class has been specified. Then, utilizing Bangsø’s plug and play techniques, whenever an instance of this class is created, the junction tree for that class is plugged into the model and appropriate evidence is applied.

Such techniques provide two main benefits. Firstly, the original structure of the network is maintained and secondly, efficiency is gained during any modifications required to the junction tree of a class by the utilization of Incremental Compilation [21] techniques.

3.3 Comparison

To put the two applications into context, let us consider the objective driving the modeling building operation in each application. SPOOK is a query driven system (i.e. the objective driving the modeling building operation is to answer a user query). SPOOK receives a query from the user and using the attributes relevant to the request, constructs a model from the knowledgebase which is reasoned on to provide an answer to the query.

Our situation assessment application is a combination of bottom up data driven and top down goal driven approaches. The users intent is used to provide context that decides which relationships are of interest in producing the situation assessments (goal driven process). The signal data is then used to determine which of these chosen relationships currently hold (evidence driven process). The objective driving the modeling building operation is to determine which of the relationship of interest to the user hold at any given time given the signal data. The overall task of producing situation assessments is constant, but the set of concepts relevant to performing this task varies dynamically. The information in the knowledgebase will flow directly from the observations and therefore much of the information in the knowledge base will be instances. In our application, we require inferencing algorithms for instances which can take advantage of the structure of the domain and reuse computations when possible.

We have implemented the SVE algorithm and are currently implementing the KBMC algorithm for OPRMs in order to run some performance comparison experiments against our implemented plug and play inference algorithms. We are also developing an enhanced SVE algorithm that will test for those cases where encapsulation doesn’t hold in the presence of instance statements. This algorithm will use SVE for classes and those instances where encapsulation holds and KBMC in those cases where it does not.

We are also currently investigating the incorporation of identity uncertainty into the language by adding an equivalence relation definition to the relational component of the OPRM definition. Like most current FOPL approaches, OPRMs currently employ the unique names assumption. That is, each instance in the knowledge base is assumed to correspond to a different object. This assumption may be violated in domains where there is a distinct possibility that multiple observations (and therefore multiple instances in the knowledge base) may represent the same object. This type of uncertainty would have a profound impact on data association. A recent thesis by [14] investigated the incorporation of identity uncertainty into PRMs.

4 Future Work

We are investigation the extension of OPRMs to dynamic domains. The expressive power of OPRMs makes it easy to construct models whose equivalent OOBN has very large cliques. The incorporation of identity uncertainty and the extension to dynamic domains will only exacerbate this problem. So we are exploring suitable appropriate approximate inference algorithms. Following the concept that OPRMs are translated into the 'equivalent' OOBN, dynamic OPRMs (DOPRMs) would be unrolled into the 'equivalent' dynamic bayesian networks (DBNs), so any of the approximate inference techniques used for DBNs could potentially be useful. We are, however, hoping to retain the benefit afforded by the pre-specification of class interfaces. Suitable approximate inferencing techniques may include Rao-Blackwellised Particle Filters, Factored Particle Filtering (FP) and Sample Propagation.

PRMS were recently extended to dynamic domains [22]. Dynamic PRMs (DPRMs) construct a DBN where each time slice and its dependence on the previous timeslice are represented by a PRM. An inferencing technique based on Rao-Blackwellized Particle Filtering was presented. However, this technique relied on two very restrictive assumptions: firstly, that reference attributes with uncertain values cannot be parents of unobserved attributes in the DPRM (i.e. violates R3.3.2) and secondly that each reference attribute refers to only one object (i.e. violates R4). These assumptions limit the ability of DPRMs to represent uncertainty over the structure of the model.

5 Conclusions

As relational databases are a common mechanism for representing structured data (e.g. medical records, sales and marketing information, etc), OPRMs are applicable to a wide range of domains and applications for example, disaster management and computer network security and stock market modeling.

In this paper, we have compared inferencing techniques for PRMs and OPRMs. Based on our current implementation of the SVE algorithm, we have discussed possible extensions such as explicit encapsulation checking and the incorporation of identity uncertainty. The first extension enhances performance; the second significantly expands the modeling capability compared to the situation assessment models produced by Pfeffer.

References

1. Lambert, D.A. Grand Challenges of Information Fusion. in Proceedings of the Sixth International Conference on Information Fusion. 2003. Cairns, Queensland.
2. Steinberg, A.N., C.L. Bowman, and F.E. White. Revisions to the JDL Data Fusion Model. in The Joint NATO/IRIS Conference. 1998. Quebec, Canada.
3. Howard, C. and M. Stumptner. Situation Assessment with Object Oriented Probabilistic Relational Models. in Proceedings of the Seventh International Conference on Enterprise Information Systems (to appear). 2005. Miami.
4. Wright, E., et al. Multi-Entity Bayesian Networks for Situation Assessment. in Proceedings of the Fifth International Conference on Information Fusion. 2002.

5. Laskey, K.B. and S.M. Mahoney. Network Fragments: Representing Knowledge for Constructing Probabilistic Models. in Proceedings of the Thirteenth Annual Conference on Uncertainty in Artifical Intelligence (UAI-97). 1997. Providence, Rhode Island: Morgan Kaufmann.
6. Das, S., R. Grey, and P. Gonsalves. Situation Assessment via Bayesian Belief Networks. in Proceedings of the Fifth International Conference on Information Fusion. 2002. Annapolis, MD, USA.
7. Blandon, P., R.J. Hall, and W.A. Wright. Situation Assessment Using Graphical Models. in Proceedings of the Fifth International Conference on Information Fusion. 2002. Annapolis, MD, USA.
8. Sutton, C., et al. A Bayesian Blackboard for Information Fusion. in Proceedings of the Seventh International Conference on Information Fusion. 2004. Stockholm, Sweden.
9. Suzic, R. Representation and Recognition of Uncertain Enemy Policies Using Statistical Models. in In Proceedings of the NATO RTO Symposium on Military Data and Information Fusion. 2003. Prague, Czech Republic.
10. Okello, N. and G. Thoms. Threat Assessment Using Bayesian Networks. in Proceedings of the Sixth International Conference on Information Fusion. 2003. Cairns, Queensland.
11. Pfeffer, A.J., Probabilistic Reasoning for Complex Systems, PhD thesis in Department of Computer Science. 1999, Stanford University. p. 304.
12. Koller, D. and A. Pfeffer. Probabilistic Frame-Based Systems. in Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98). 1998. Madison, Wisconsin.
13. Getoor, L., Learning Statistical Models From Relational Data, PhD thesis in Department of Computer Science. 2002, Stanford University. p. 189.
14. Pasula, H.M., Identity Uncertainty. 2003: PhD thesis in Department of Computer Science, University of California, Berkeley.
15. Bangso, O., Object Oriented Bayesian Networks, PhD thesis in Department of Computer Science. 2004, Aalborg University. p. 110.
16. Koller, D. and A. Pfeffer. Object-Oriented Bayesian Networks. in Proceedings of the Thirteenth Annual Conference on Uncertainty in Artifical Intelligence (UAI-97). 1997. Providence, Rhode Island: Morgan Kaufmann.
17. Bangso, O., M.J. Flores, and F.V. Jensen, Plug and Play Object Oriented Bayesian Networks. Lecture Notes in Artificial Intelligence. Vol. 3040. 2004.
18. Bangso, O. and P.-H. Wuillemin, Object Oriented Bayesian Networks. A Framework for Top Down Specification of Large Bayesian Networks with Repetitive Structures. 2000, Technical Report CIT-87.2-00-obphw1, Department of Computer Science, Aalborg University.
19. Bangso, O. and P.-H. Wuillemin. Top Down Construction and Repetitive Structures Representation in Bayesian Networks. in Proceedings of the Thirteenth International Florida Artificial Intelligence Research Society Conference. 2000: AAAI Press.
20. Pfeffer, A., et al. SPOOK: A System for Probabilistic Object Oriented Knowledge Representation. in Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI-99). 1999.
21. Flores, M.J., J.A. Gamez, and K.G. Olsen. Incremental Compilation of a Bayesian Network. in Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence. 2003: Morgan Kaufmann.
22. Sanghai, S., P. Domingos, and D. Weld. Dynamic Probabilistic Relational Models. in Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence. 2003. Mexico.

A Network Service Access Control Framework Based on Network Blocking Algorithm*

Jahwan Koo[1] and Seongjin Ahn[2,**]

[1] School of Information and Communications Engineering, Sungkyunkwan Univ.
Chunchun-dong 300, Jangan-gu, Suwon, Kyounggi-do, Korea
jhkoo@songgang.skku.ac.kr

[2] Department of Computer Education, Sungkyunkwan Univ.
Myeongnyun-dong 3-ga 53, Jongno-gu, Seoul, Korea
sjahn@comedu.skku.ac.kr

Abstract. Currently, the major focus on the network security is securing individual components as well as preventing unauthorized access to network service. In this paper, we propose a network blocking algorithm and architecture, which provides a comprehensive, host-by-host perspective on IP over Ethernet network security. The proposed system is also an effective tool for managing network resources containing IP address, MAC address and hostname, etc. under diverse and complicated network environment. Therefore, we will be able to enhance local network security on the wired and wireless environment with the network resource and security management system based on our proposed framework.

1 Introduction

Delivery of a packet to a host or router requires two levels of addresses. One is a logical address(IP address) at the network layer and the other is a physical address(also is known hardware address, network adapter address, or MAC) at the link layer. The Internet Protocol(IP), however, was designed to be independent of any specific link layer. Therefore, there is no way to determine the destination physical address from the logical address. Because there is no correlation between these two addresses, it is impossible to derive one address from the other. Hence, we need to be able to map a logical address to its corresponding physical address and vice versa. These can be done using either static or dynamic mapping. Static mapping means creating a table that associates a logical address with a physical address. This table is statically stored in each machine on the network. Every time a physical address is changed, the table must be updated. This method creates a huge overhead. As usual, the mapping can be done dynamically, which means that the sender asks the receiver to announce its physical address when needed. The address resolution protocol(ARP) is a dynamic mapping method that finds a physical address given a logical address and a basic protocol in every Internet host and router.

* This paper was supported by 63 Research Fund, Sungkyunkwan University, 2004

** Dr. S. Ahn. is the Corresponding Author

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 54–61, 2005.

It is true that the ARP has many security weaknesses. Although there are many papers related the ARP vulnerability, most of them are mainly focused to security problems, a variety of attacks based on these problems, including ARP sniffing, DoS, hijacking, and cloning, and defenses against these attacks[1]. In this paper, however, we propose a network blocking algorithm and architecture for network resource and security management system using ARP security weaknesses conversely.

The rest of the paper is organized as follows. In section 2, we describe comprehensive surveys including the basic operation of ARP and the mechanism of ARP cache, which are based on our network blocking algorithm. In section 3, we propose a network blocking algorithm and architecture for network resource and security management. The final section offers some concluding remarks.

2 Address Resolution Protocol

2.1 Overview of ARP

As RFC 826[2] describes, ARP is the protocol used by shared access, broadcast-based networking technologies such as Ethernet and Token Ring. This protocol is used to resolve the next-hop IP address of a node to its corresponding hardware address.

Let us see how ARP operations on a typical Internet. For convenience, we denote the structure of the ARP frame including Ethernet header as [Destination, Source, (Operation, Sender Hardware Address, Sender Protocol Address, Target Hardware Address, Target Protocol Address)] in short. These are the steps involved in an ARP process:

1. The sending node wants to send a IP datagram to another node and it knows the IP address of that node.
2. The sending node always examines the content of ARP cache(to be discussed in section 2.2) before an ARP request message is sent. If its ARP cache table has an entry for the destination node, this task is easy(go to step 8). Otherwise, go to next step.
3. If the ARP cache table does not find an entry for the destination node, the sending node asks ARP to create an ARP request message, filling in the sending physical address, the sending IP address, and the destination IP address. The destination physical address field is filled with 0s.
4. The message is passed to the link layer where it is encapsulated in a frame using the physical address of the sending node as the source address and the physical broadcast address as the destination address. In other word, the ARP request frame is [ff:ff, SHA, (1, SHA, SPA, 0, TPA)].
5. Every host or router on the LAN receives the frame. All the other nodes except the destination drop the packet. The destination node recognizes the IP address.
6. The destination node responds with an ARP reply message that contains its physical address. The message is unicast. In other word, the ARP reply frame is [SHA, DHA, (2, DHA, DPA, SHA, SPA)].

7. The sending node receives the ARP reply message. When an ARP exchange is complete, both the sending node and the destination node have each other's IP address-to-MAC address mapping in their ARP caches.
8. The sending node knows the physical address of the destination node using the ARP cache entry's lookup.
9. The IP datagram, which carries data for the destination node, is now encapsulated in a frame and is sent to the destination.

The purpose of the ARP exchange is to query all the other nodes on the LAN to determine the physical address corresponding to the IP address that is being resolved and is to maintain the IP address-to-MAC address mapping table automatically[3].

2.2 Mechanism of ARP Cache

The maintenance of an ARP cache on each node is essential to the efficient operation of ARP. This cache maintains the recent mappings from logical addresses to hardware addresses in a RAM-based table. As RFC 1122[4] describes, some mechanisms (i.e. Timeout, Unicast poll, Link-layer advice, and Higher-layer advice) have been used, sometimes in combination, to maintain out-of-date cache entries.

- **When does an entry be created in the ARP cache?** When transferring IP packets from the sending node to the destination node, an entry in both node's ARP cache is created through the ARP request-reply message exchange. In the sending node's ARP cache, the destination's IP-to-MAC is created, and in the destination's ARP cache, the sending node's IP-to-MAC is created. Also, It has been proved, needless to say, that the ARP cache entry can be created by the network browsing service, NetBIOS communication, or network file and printer sharing as well as the ARP request-reply exchange in the Window environment.
- **When does an entry be deleted in the ARP cache?** The ARP cache entry's timeout works differently on the TCP/IP implementations. Berkeley-derived implementations normally have a timeout of 20 minutes for a completed entry(both the ARP request and the ARP reply are sent and received) and 3 minutes for an incomplete entry(where an ARP request is sent to a nonexistent host on the LAN)[5]. Windows Server 2003 family and Windows XP persist for only 2 minutes. If the ARP cache entry is not used within 2 minutes, the entry is deleted; otherwise, if it is used within 2 minutes, it is given additional time in 2-minute increments, up to a maximum lifetime of 10 minutes. After a maximum of 10 minutes, the ARP cache entry is removed and must be resolved through another ARP exchange[6]. In Linux, if the ARP cache entry is not used within 2 minutes, the entry is deleted; otherwise, if it is used within 2 minutes, the timeout value is restart. If the ARP cache reaches its maximum capacity before entries expire, the oldest entry is deleted so that a new entry can be added.

- **When does an entry be updated in the ARP cache?** Windows Server 2003 family, Windows 2000 and XP are to update the ARP cache entry with additional time, in 2-minute increments, while it is in use. Also, they are updating the ARP cache entry when the receipt of an ARP request sent by the node with the ARP cache entry's IP address. When an ARP request that was sent by an IP node corresponding to an existing entry in the ARP cache is received, the ARP cache entry is updated with the received ARP request's MAC address.

3 Network Service Access Control Framework

In this section, we describe a network service access control framework based on the network blocking algorithm. We sometimes have some experiences on the IP address conflict and network inaccessibility. Without administrator's admission, the unauthorized user abuses the network configuration resources like IP address, MAC address, and hostname on the TCP/IP network environment. It is basically caused by the security weakness of TCP/IP protocols and the deficiency of network administration. Therefore, there is urgent need to solve the problems from the viewpoint of network management and operation. The architecture for the network resource and security management can be used to block the network accessibility of unauthorized users and isolate them from authorized users on the same network management domain.

It is typically based on a manager-to-probe or a manager-to-agent model. At least one node on the entire network is designated as a manager system and a probe or an agent system has to be installed on each network management domain.

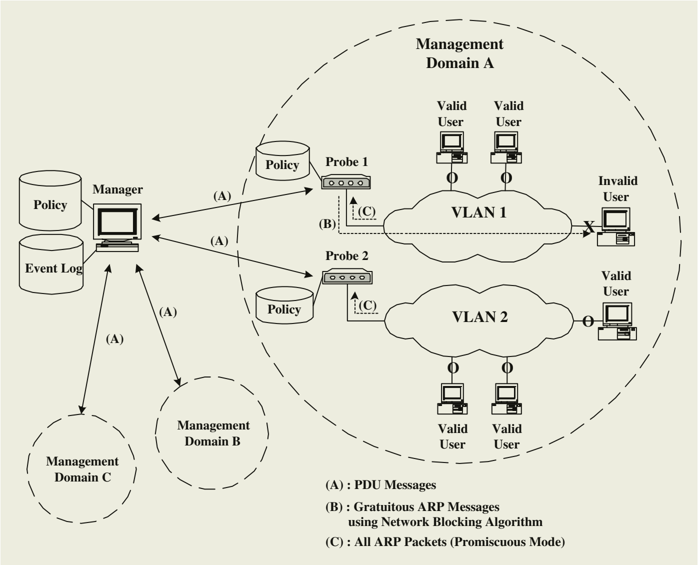


Fig. 1. A Network Service Access Framework

3.1 Overall Architecture

Figure 2 and 3 show the proposed architecture for the network resource and security management. Each network management domain contains three major components: manager, probe, and management policy.

- **Manager.** At least one host on the network is designated as a network control host, or manager. The manager serves as the interface for the human network manager into the network resource and security management system. It will have a set of management applications for management, report, communication, and database. And, it will have the capability of translating the network manager's requirements into the actual monitoring and control of remote probes in the network. Also, it will maintain the network resource management information from the probes via the PDU message exchanges, visualize current resource status information such as the number of total IP addresses, used IP addresses, and unused IP addresses and provide the real-time information such as the corresponding IP-to-MAC addresses, hostnames, and policies.

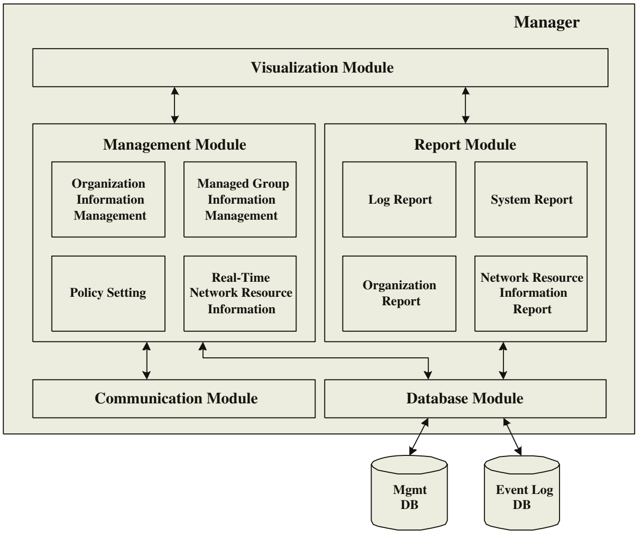


Fig. 2. Proposed architecture for manager system

- **Probe.** A Probe will be installed on each network management domain. It responds to requests for information and actions from the manager, collects all ARP messages at the promiscuous mode and examines them whether they are abnormal on the basis of the management policy to be defined. If so, it sends gratuitous ARP packets via the network blocking algorithm to all the nodes of the corresponding domain and updates properly the ARP cache entries of managed nodes.

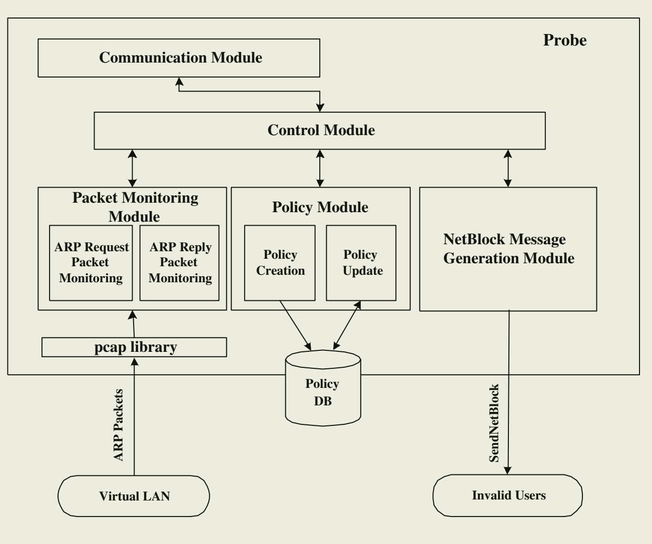


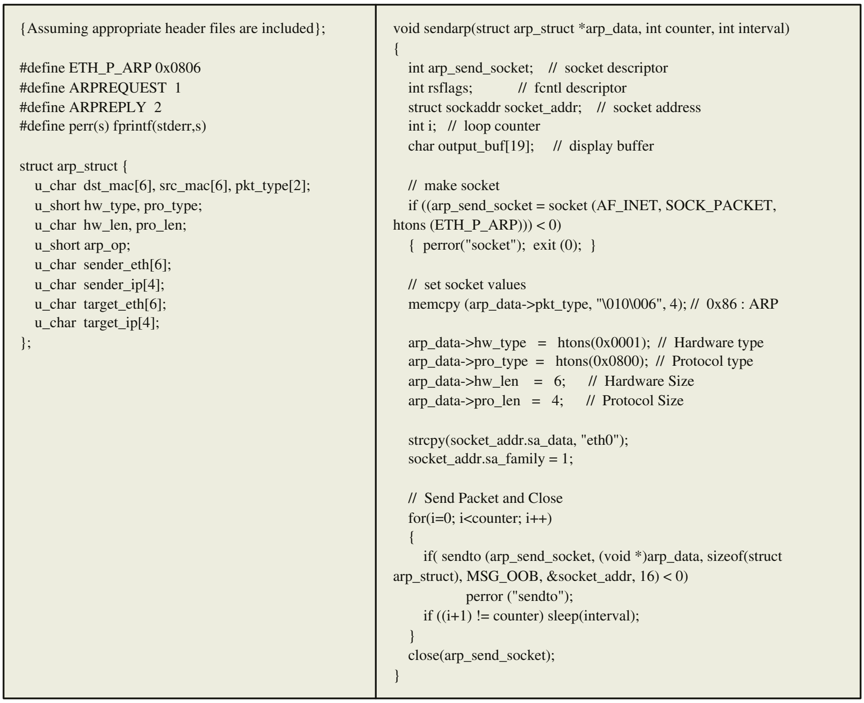
Fig. 3. Proposed architecture for probe system

- **Management Policy.** Network administrator can set policies that enables optimized IP address management. Management Policy is divided into six: IP blocking, MAC blocking, IP/MAC unblocking, IP static, MAC static, and No policy. IP blocking policy restraints the network accessibility of nodes with unauthorized IP address, MAC blocking policy restraints the network accessibility of nodes with unauthorized MAC address, IP static policy ties the corresponding IP address to the specific MAC address, MAC static policy ties the corresponding MAC address to the specific IP address and No policy does not manage anything. For example, by banning use of IP addresses and MAC addresses by unauthorized users, the use of IP address can be regulated and network can be protected.

3.2 Network Blocking Algorithm

Network blocking algorithm broadcasts ARP messages repeatedly to all the nodes on the management domain containing the specific node.

The *sendarp* function takes a pointer to the data structure of ARP, forms a gratuitous ARP packet for the specific node's IP address, and transmits it recursively to the target node. We implement simple send-arp command using this algorithm. The syntax of this command is send-arp *filename number interval-time*, where *filename* is an ASCII file containing ARP frame information such as destination, source, operation, SHA, SPA, THA, and TPA, *number* is the number of ARP packet that will be send, and *interval-time* is the interval time between the moment an ARP packet sends and the moment next ARP packet sends. This command sends out an ARP packet with source-to-target IP and MAC addresses supplied by the administrator.



```
{Assuming appropriate header files are included};

#define ETH_P_ARP 0x0806
#define ARPREQUEST  1
#define ARPREPLY  2
#define perr(s) fprintf(stderr,s)

struct arp_struct {
   u_char  dst_mac[6], src_mac[6], pkt_type[2];
   u_short hw_type, pro_type;
   u_char  hw_len, pro_len;
   u_short arp_op;
   u_char  sender_eth[6];
   u_char  sender_ip[4];
   u_char  target_eth[6];
   u_char  target_ip[4];
};
```

```
void sendarp(struct arp_struct *arp_data, int counter, int interval)
{
   int arp_send_socket;   //  socket descriptor
   int rsflags;           //  fcntl descriptor
   struct sockaddr socket_addr;   //  socket address
   int i;   //  loop counter
   char output_buf[19];     //  display buffer

   //  make socket
   if ((arp_send_socket = socket (AF_INET, SOCK_PACKET,
htons (ETH_P_ARP))) < 0)
   {  perror("socket");  exit (0);  }

   //  set socket values
   memcpy (arp_data->pkt_type, "\010\006", 4); //  0x86 : ARP

   arp_data->hw_type   =   htons(0x0001);  //  Hardware type
   arp_data->pro_type  =   htons(0x0800);  //  Protocol type
   arp_data->hw_len    =   6;      //  Hardware Size
   arp_data->pro_len   =   4;      //  Protocol Size

   strcpy(socket_addr.sa_data, "eth0");
   socket_addr.sa_family = 1;

   //  Send Packet and Close
   for(i=0; i<counter; i++)
   {
      if( sendto (arp_send_socket, (void *)arp_data, sizeof(struct
arp_struct), MSG_OOB, &socket_addr, 16) < 0)
               perror ("sendto");
      if ((i+1) != counter) sleep(interval);
   }
   close(arp_send_socket);
}
```

Fig. 4. Network blocking message generation module in C language

4 Conclusions

In this paper, we presented comprehensive surveys and experimental results of the basic operation of ARP and the mechanism of ARP cache. Their purpose is to propose a network blocking algorithm and architecture. Sometimes network resources including IP address, MAC address, and hostname could be misused for the weakness of TCP/IP protocol suite and the deficiency of network management. Therefore, we proposed a network service access control framework based on the network blocking algorithm and architecture. The basic concept of the proposed network resource and security management system is that authorized users can access their own network but unauthorized users should not be able to access. The proposed system is an effective tool for managing network resources containing IP address, MAC address and hostname, etc. under diverse and complicated network environment. This system collects the information of network resources, prevents specific IP addresses' conflicts, blocks the accessibility of unauthorized users, and redistributes available IP addresses. Therefore, network resources should be managed by the life cycles of IP address generation, usage, and revocation on the basis of the specific policies such as IP/MAC blocking/unblocking and IP/MAC static.

References

1. S.M. Bellovin. “Security Problems in the TCP/IP Protocol Suite,” Computer Communication Review, Vol. 19, No. 2, pp 32-48, April 1989.
2. David C. Plummer. “An Ethernet Address Resolution Protocol,” RFC 826, November 1982.
3. Comer, Douglas. “Internetworking with TCP/IP: Vol. 1, Principles, Protocols, and Architecture,” 4th edition, Prentice Hall, 2000.
4. R. Braden. “Requirements for Internet Hosts – Communication Layers,” RFC 1122, October 1989.
5. Stevens, W. Richard. “TCP/IP Illustrated. Volume 1: The Protocols. Reading,” Addison-Wesley, 1993.
6. J. Davies and T. Lee. “Microsoft Windows Server 2003 TCP/IP Protocols and Services Technical Reference,” Microsoft Press, 2003.
7. M. Maxim and D. Pollino. “Wireless Security,” RSA Press, 2002.

3-D Building Reconstruction Using IKONOS Multispectral Stereo Images

Hong-Gyoo Sohn, Choung-Hwan Park, and Joon Heo

School of Civil Engineering, Yonsei University, Seoul, Korea
{sohn1,c142520,jheo}@yonsei.ac.kr

Abstract. This paper presents an effective strategy to extract the buildings and to reconstruct 3-D buildings using high-resolution multispectral stereo satellite images. Proposed scheme containes three major steps: building enhancement and segmentation using both Background Discriminant Transformation (BDT) and ISODATA algorithm, conjugate building identification using the object matching with Hausdorff distance and color indexing, and 3-D building reconstruction using photogrammetric techniques. IKONOS multispectral stereo images were used to evaluate the scheme. As a result, the BDT technique was verified as an effective tool for enhancing building areas since BDT suppressed the dominance of background to enhance the building as a non-background. In building recognition, color information itself was not enough to identify the conjugate building pairs since most buildings are composed of similar materials such as concrete. When both Hausdorff distance for edge information and color indexing for color information were combined, all segmented buildings in the stereo images were correctly identified. Finally, 3-D building models were successfully generated using the space intersection by the forward Rational Function Model (RFM).

1 Introduction

3-D building reconstruction in urban areas is one of the highlighted issues in photogrammetry and remote sensing. Generated 3-D building information can be used in various fields such as urban planning, disaster management, navigation system, and cyber city. In order to extract 3-D building information, various attempts have been performed using aerial images, satellite images, LIDAR data, Digital Surface Model (DSM), Digital Elevation Model (DEM), and Geospatial Information System (GIS) thematic maps. Aerial and satellite images are the primary data sets used in most researches.

Conventional 3-D building reconstruction techniques are divided into image-based and model-based approaches. Image-based approach utilizes all possible extracted data from the images. Several attempts to extract and construct 3-D building models using the aerial imagery were investigated [6], [8]. These trials were failed to match the extracted buildings simultaneously since only edge information is not enough for successful matching. Subsequently, a system that detects and constructs 3-D models from a single image was developed [5]. In this case, shadow information was used to support the building hypotheses. However, shadow is not always available information. Although approaches using color information in satellite imagery were also performed [7], extraction method was limited only to color information.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 62–68, 2005.

In model-based approach, buildings in the image are extracted using the prior building models. Therefore, model-based approach still requires interacting with the system and input information such as the approximate building position, the initial building model, and model parameters. Fischer and others [2] solved the building extraction problem using semi-automated approach. This has a deficiency in that users manually define the building model and find the building elements in one image.

This paper presents an effective strategy to extract building and to establish 3-D building models using high-resolution multispectral stereo satellite images. Proposed scheme consists of three major processes.

2 Building Extraction and Recognition

The detection and extraction of objects in images is often dependent on the suppression of the background. Also, this problem is of great importance in pattern recognition systems. The success of automated pattern recognition systems depends on the enhancement of significant features in relation to the irrelevant background information in the pre-processing stage.

2.1 Background Discriminant Transformation (BDT)

A large variety of techniques are in use for enhancing significant features in images. The commonly used techniques are linear and non-linear stretching, histogram equalization, spatial filtering in univariate space, and linear transformation in multivariate space [9]. The BDT is one of the linear transformations for image enhancement. The BDT is designed to discriminate between the background and the non-background (for example, buildings of interest). One of the importance aspects of this technique is that it is scale-invariant.

In the BDT technique, the image is assumed to have two classes: background and non-background. In order to enhance the non-background class, the axes in spectral space are to be rotated to reduce the background variability and to increase the non-background variability, much like in the Principal Component Transformation (PCT). In other words, BDT coefficients are computed to maximize the variance of the non-background relative to the background. The theoretical basis of this algorithm is well described in Carroll and others [1].

The BDT computes the same number of new calculated bands as the original multispectral bands. First band means that the ratio of variances of non-background class and background class is the maximum. This band shows the dominance of non-background over the background. Last band means that the ratio of variances is the minimum, and thus background class is dominant over the non-background class. After enhancing the original multispectral image using BDT, it is possible to segment the image into a limited number of clusters corresponding to non-background and background. In this study, the BDT is used to enhance building areas and ISODATA algorithm is used to segment buildings from the enhanced images.

2.2 Hausdorff Distance and Color Indexing

The operation of object matching consists of deciding if two objects observed in different scenes are identical. Object matching often plays important role in object recognition. In this study, the Hausdorff distance and color indexing value are used to compute matching scores in object matching.

Given two finite point sets $A = \{a_1, \cdots, a_p\}$ and $B = \{b_1, \cdots, b_q\}$, the Hausdorff distance is defined as [4]:

$$D(A, B) = \max(d(A, B), d(B, A)) \tag{1}$$

where $d(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\|$ and $\| \cdot \|$ is Euclidean norm on the points of A and B. The function $d(A, B)$ is called the direct Hausdorff distance from A to B. The distance $d(A, B) \geq 0$ and if $d(A, B) = 0$, the two data sets are identical. The distance $d(A, B)$ depends on the relative translation, rotation, and scale change between two data sets. The segmented objects of interest do not necessarily have the same position, orientation, and scale in each data set. Therefore, the two objects need to be registered before the distance between them can be computed. Final edge matching score between the two edge images E_1 and E_2 is then rewritten as:

$$D_{new} = D(E_1^T, E_2) \tag{2}$$

where D_{new} is new Hausdorff distance between E_1^T and E_2.

Color indexing algorithm identifies an object by comparing its colors with the colors in other image. However, crucial point for this processing is that the total area covered by each color must be taken into account. The areas are computed and compared by histogramming the images and intersecting the histograms. A color histogram is three dimensional and simply represents the count of the number of pixels in image having a particular RGB value. Color histograms of images of every object are computed and stored. Presented with image of an unknown object, the color indexing algorithm computes its color histogram and intersects it with every one of the stored histograms in order to find the one that matched best.

For two objects with color histograms, the color matching score is defined as [3]:

$$C = \frac{\sum_r \sum_g \sum_b \min(H_1(R, G, B), H_2(R, G, B))}{\min(|H_1|, |H_2|)} \tag{3}$$

where $H_i(R, G, B)$ is the color histogram of i th object. $0 \leq C \leq 1$ and if $C = 1$, one of the histograms is totally included in the other one.

3 3-D Building Reconstruction

In order to generate 3-D building model, 3-D position information are essential. 3-D position of buildings can be calculated using photogrammetric techniques. 3-D positioning is performed with conjugate points in stereo images and geometric model expressed relationships between image space and ground space.

For acquiring accurate conjugate points, least squares image matching technique is introduced. Least squares image matching is applied along the edges of identified building pairs. For calculating 3-D position of buildings, forward Rational Function Model (RFM) is used. RFM, one of the replacement sensor models, has been extensively used in IKONOS imagery. The RFM defines the relation between the image space and ground space in the form of polynomial ratios.

$$r=\frac{p_1(\varphi,\lambda,h)}{p_2(\varphi,\lambda,h)},\quad c=\frac{p_3(\varphi,\lambda,h)}{p_4(\varphi,\lambda,h)} \tag{4}$$

where r and c are the normalized row and column indices of pixels in the image space and φ , λ , and h are the normalized coordinates in the ground space. The detailed algorithm about space intersection by RFM is presented in Sohn and others [10].

4 Experiments and Results

To test the proposed 3-D building reconstruction scheme, IKONOS multispectral stereo images taken in February 7, 2000 (4 bands/1 m resolution) were used. The stereo pairs cover the San Diego City, U.S.A. as shown in Figure 1. The test area highlighted in Figure 1 was selected for 3-D building reconstruction. Unlike conventional satellite, which takes cross-track stereo images from different orbital passes, IKONOS collects same pass stereo pairs. That is the two images constituting the stereo pair are taken on the same orbital pass. Stereo pairs used in this study are scanned at reverse direction. The nominal elevation angle of each image is 62.1° and 64.6°. Also, it is appropriate to test proposed algorithms since stereo images contain various man-made features and natural topography.

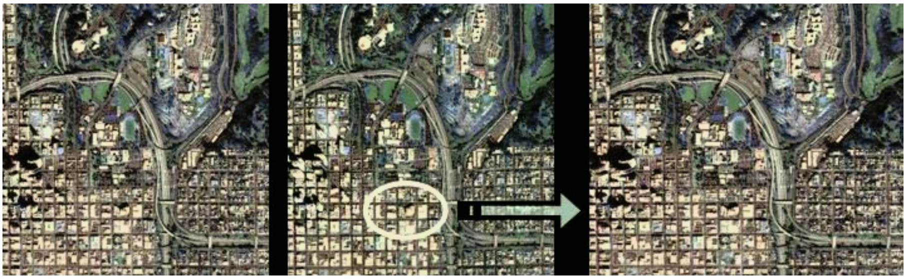


Fig. 1. IKONOS multispectral stereo images and test area

For building enhancement and segmentation, five training sites for background class are collected in stereo images. Mean vector and covariance matrices for background class and those of whole image are calculated. After enhancing images using the BDT technique, buildings are segmented using ISODATA algorithm. Figure 2 shows extracted buildings in the stereo images. Total 34 buildings in left image are extracted, while 36 buildings in right image are extracted. However, a small group of buildings are not extracted since some buildings are obscured by the shadow of neighboring buildings.

Extracted buildings first have to be identified as conjugate building pairs before performing point matching for accurate conjugate points. Especially, building recognition is effective when extracted buildings have not same number in each image. Object matching algorithm using Hausdorff and color indexing technique is used for building recognition. From a practical point of view, this approach has an advantage that it does not need an additional process such as epipolar image resampling.

To combine the color matching score and edge matching score into a single matching score, it should be noted that the color matching score is a similarity index while the edge matching score is dissimilarity index.

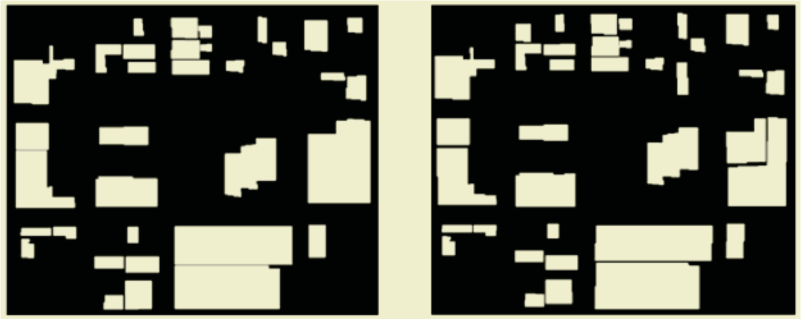


Fig. 2. Extracted buildings in test area (left: 34 buildings. right: 36 buildings)

Therefore, we have chosen the following equation to compute the total matching score as follows:

$$T = P^{m}_{color} \cdot C - P^{m}_{edge} \cdot D \qquad (6)$$

where T is the total matching score, C and D are color indexing score and Hausdorff distance, and P^{m}_{color} and P^{m}_{edge} are ratios of successful matched objects over the candidate matching objects.

Table 1 summarizes the results of building recognition when both matching scores are combined. R^{a} is the average rank in actual pairs, R^{m} is the maximum rank in actual pairs, M^{t} (or M^{f}) is the number of true (or false) matches.

Table 1. Building recognition results using the object matching technique

	R^{a}	R^{m}	M^{t}	M^{f}	P^{m} (%)
Edge	1.1	2	32	2	94.1
Color	19.8	32	4	30	11.8
Edge+Color	1.0	1	34	0	100

As shown in Table 1, the accuracy of building recognition using only color information is considerably low. The result may be caused by the similar reflectance characteristics of the building. In case when buildings are composed of materials such as concrete, they generally have the homogeneous spectral characteristics. However, all buildings are correctly recognized when both edge information and color information are combined.

Once the initial position for point matching is defined by building recognition process, least squares image matching is further performed to obtain the accurate conjugate points. Total 34 corresponding building pairs in stereo images were successfully matched. The final 34 building pairs, which contain 4,892 conjugate points, were used to calculate 3-D position by the forward RFM. Calculated 3-D coordinates have geodetic coordinates on WGS-84 ellipsoid. Therefore, additional coordinate conversion to UTM is followed. Final 3-D building reconstruction results are overlaid on geocoded reference image as shown in Figure 3. In Figure 3, the calculated heights of buildings were three-times vertically exaggerated.

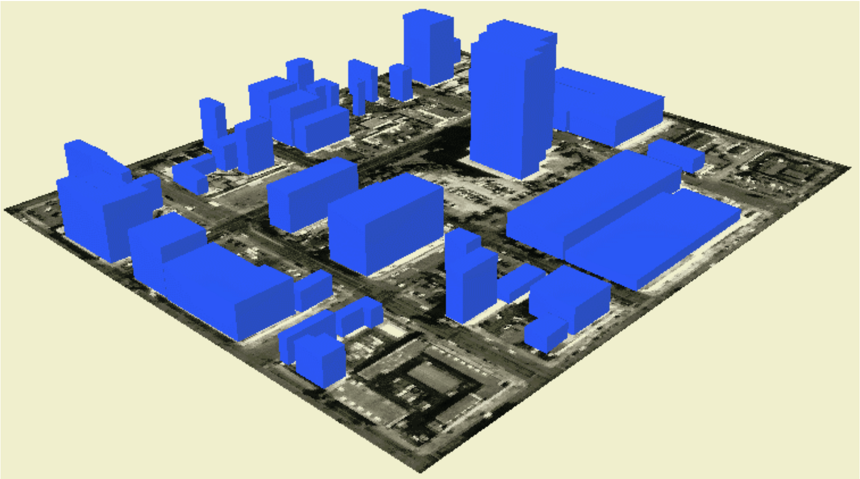


Fig. 3. 3-D building reconstruction results of the test area

5 Conclusions

Automatic or semiautomatic techniques for building extraction and 3-D building reconstruction have evolved recently. They showed great potential for generating 3-D building models. In this paper, we presented a new method for 3-D building reconstruction using high-resolution multispectral stereo images. Our scheme focused on finding solution about following question: the effect of color information in 3-D building reconstruction. As a result, color information provided a useful tool for building extraction. Background and non-background such as building can be surely separated in multispectral imagery. Especially, the BDT technique based on spectral characteristics of objects is suitable for multispectral imagery. However, color information in building recognition did not provide satisfactory result since buildings are often composed of almost homogeneous materials, which causes similar spectral characteristics in resulting images. It was also confirmed that IKONOS stereo images, which adopt replacement sensor model as a basic sensor model, are suitable for the 3-D building reconstruction. Our approach can be extended to generate 3-D building model using other multi-source data sets.

References

1. Carroll, J. D., Green, P. E., and Chaturvedi, A.: Mathematical tools for applied multivariate analysis, Elsevier Science, Revised edition, (1997) 259-294
2. Fischer, A., Kolbe, T. H., and Lang, F.: On the use of geometric and semantic models for component-based building reconstruction proceedings, SMATI (Semantic Modeling for the Acquisition of Topographic Information from Images and Maps) Workshop, (1999) 101-119
3. Funt, B. V., Finlayson, G. D.: Color constant color indexing, IEEE Transaction on Pattern Analysis and Machine Intelligence, 17(5), (1995) 522-529
4. Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J.: Comparing images using the hausdorff distance, IEEE Transaction on Pattern Analysis and Machine Intelligence, 15(9), (1993) 850-863
5. Lin, C., Nevatia, R.: Building detection and description from a single intensity image, Computer Vision and Image Understanding, 72(2), (1998) 101-121
6. Noronha, S., Nevatia, R.: Detection and description of buildings from multiple aerial images, IEEE Transaction on Pattern Analysis and Machine Intelligence, 23(5), (2001) 501-518
7. Ormsby, J. P.: Evaluation of natural and man-made features using Landsat TM data, International Journal of Remote Sensing, 13, (1992) 303-318
8. Roux, M., McKeown, D. M.: Feature matching for building extraction from multiple view, IEEE Proceeding of Computer Vision and Pattern Recognition, (1994) 46-53
9. Shettigara, V. K.: Image enhancement using background discriminant transformation, International Journal of Remote Sensing, 12(10), (1991) 2153-2167
10. Sohn, H.-G., Park, C.-H., and Chang, H.: RFM-based image matching technique using object-space approach, Journal of Japan Society of Photogrammetry and Remote Sensing, 43(4), (2004) 4-12.

Heuristic Algorithm for Estimating Travel Speed in Traffic Signalized Networks

Hyung Jin Kim[1], Bongsoo Son[1], Soobeom Lee[2], and Sei-Chang Oh[3]

[1] Dept. of Urban Planning and Eng., Yonsei Univ., Seoul, Korea
{hyungkim,sbs}@yonsei.ac.kr
[2] Dept. of Transportation Eng., Univ. of Seoul, Seoul, Korea
mendota@uos.ac.kr
[3] Div. of Environmental, Civil and Transportation Eng. Ajou Univ., Korea
scoh@ajou.ac.kr

Abstract. Theoretic methods for travel speed estimation are common in the traffic literature, but treatments of how to account delay caused by traffic signal are less common. This paper is an attempt to improve conventional technique for evaluating travel speed in signalized intersection networks. To do so, this paper employs new concept which is associated with the vehicles' trajectory in the time-space diagram. The most useful advantage of the algorithm proposed in this paper is that, the traffic analysts can reduce much tedious works for screening and selecting a bunch of speed data.

1 Introduction

Lee *et al.* [1] has proposed a method for classifying traffic conditions with occupancy data collected from loop detector. However, this method requires a complex thing that is to define threshold value of percent occupancy time for classifying the traffic conditions. More conventionally, travel speed is typically used for classifying the traffic conditions of traffic signalized intersection networks. Theoretic methods for travel speed estimation are common in the traffic literature, but treatments of how to account delay caused by traffic signal are less common [2]. Conventional method is to estimate the average speed of all vehicles traveled during the same evaluation time period [3].

It should be noted first that a number of delays such as stopped delay, approach delay, travel-time delay and time-in-queue delay may occurred at a signalized intersection during the same time period depending upon the coordination of green times of traffic signals relatively closely spaced in roadway network. Thus, it is necessary to coordinate their green times so that vehicles may move efficiently through the set of signals. It is often called as "signal progression" in traffic engineering. However, it is almost impossible to coordinate all green times for all approaches at traffic signal intersection network. Due to this fact, some vehicles inherently experience the delays under the same traffic condition.

This paper is an attempt to improve conventional technique for evaluating travel speed in signalized intersection networks. To do so, this paper employs new concept which is associated with the vehicles' trajectory in the time-space diagram. Son *et al.* [4] have first employed the concept for predicting the bus arrival time in the signalized intersection networks.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 69–73, 2005.

2 Time-Space Diagram for Vehicles' Travel Trajectory

Figure 1 illustrates the trajectories that three vehicles take as time passes in the signalized intersection network. In the figure, trajectory type I is associated with the vehicles arrived at the traffic signal i during the red time period. Trajectory II represents the vehicles arrived at the signal during the period between ending of red time and beginning of green time and experienced delay for passing the traffic signal. Trajectory type III is related to the vehicles passed the traffic signal without any delay. The three types of trajectories indicate that the travel speeds on the signalized intersection networks are widely different and greatly dependent upon whether or not vehicles await at traffic signals. In other words, the travel speed in the signalized intersection network significantly vary depending upon the state of signals (i.e., green time or red time of the signal) as well as the coordination of green times of signals [3].

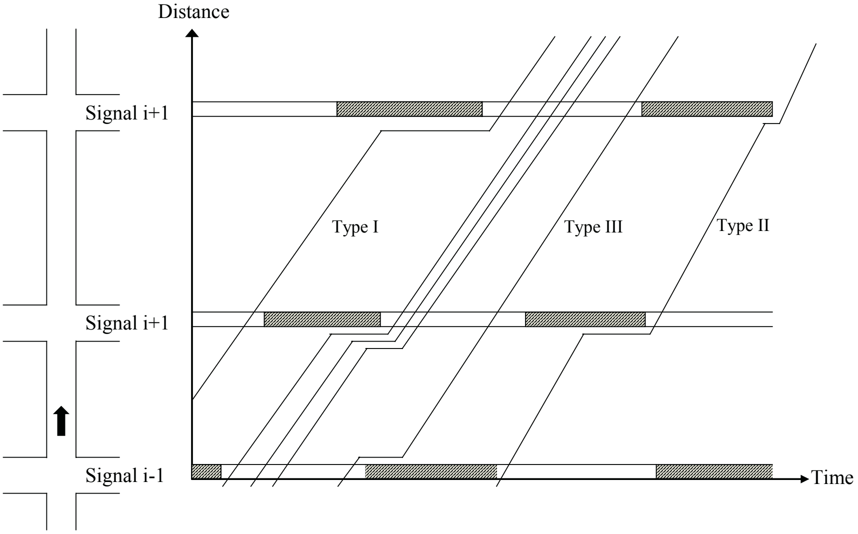


Fig. 1. Time-space diagram for vehicle trajectories in signalized network

As above-mentioned, the conventional method estimates the average speed of all vehicles passing a roadway section on signalized intersection network over some specified time period (*e.g.*, every 5 to 15 minutes), where the speed is the inverse of the time taken by the vehicles to traverse the roadway distance. However, as can be seen from Figure 1, the waiting times occurred at signalized intersections would cause major differences in estimating of each vehicle's travel speed in signalized intersection network. The conventional method definitely has limitation in reflecting the variations of vehicles' trajectory in estimating the travel speed.

3 Characteristics of Field Data

In this study, the speed data were collected by floating car method for the congested and non-congested time periods. A total of ten passenger cars were assigned to depart from the upstream of "Intersection 1" to "Intersection 5" every 1 minute time interval. The study site is 2.5 km long a major arterial that links between the western region and the old downtown area in Seoul. The study site is consisted of 6 to 8-lane sections and located near the old downtown. The speed limit of the study site is 60km/h. Table 1 summaries speed data which were calculated based on each vehicle's travel time measured from the field. In the table, the speed data were grouped for 5-minute evaluation time period.

Table 1. Travel speed data

Calculated speed values of 10 floating vehicles	
Non-congested	Congested

{45, 43}
{22, 25, 29, 35}
{29, 35, 43, 46}
	{16, 13, 12, 14, 12}
	{13, 10, 11}
	{8, 11}

The speed values measured under non-congested traffic condition were fluctuated from 24km/h to 53km/h. The fluctuation was severe during the second 5-minute evaluation time period. These results are not surprising, since the speed variation would be occurred by both the state of signals and the coordination of green times of signals. (In order to better understand some of the discussion, refer to Figure 1.)

With respect to non-congested traffic condition, the average value of speed data during the first 5-minute time period is reasonable for representing the reality, but the average values of the other two time periods are not good since the travel speeds vary even under the same traffic condition. It is logical to not wonder why travel speed of each vehicle does result in sufficient discrepancy from the average speed estimated during the same evaluation time period. If we estimate the average speed of all vehicles traveled during the same evaluation time period by using the same manner of conventional method, it may lead to misunderstanding.

More specifically, for an example, there is enough unused green time during a given green phase under non-congested traffic condition. If some vehicles arrived just after the ending of green time during the same evaluation time period, they should wait until the signal turns green again. Thus, the average travel speed would result in lower value due to the red time included during the same time period. Consequently, the average speed value may not be appropriate for representing the non-congested traffic conditions. For this case, somewhat higher speed value among all speeds measured during the same time period is more appropriate for representing the real traffic conditions rather than the average speed value. However, the data associated with the congested traffic condition, are not fluctuated so that the average speed of all vehicles traveled during the same evaluation time period seems to be reasonable for representing the reality.

4 Algorithm

The task to do in this paper is how to determine the most reasonable speed value for representing the real traffic condition. To do this, this paper proposes a heuristic algorithm for classifying traffic condition by using the speed data. The rationale of the algorithm is based on the vehicles' trajectories shown in Figure 1. That is, the maximum difference of travel time among all vehicles traveling a given distance during the same time evaluation period (or under the same traffic condition) must be, theoretically, less than the total sum of red times of all signals located on the traveling route. Otherwise, there should be non-recurrent incident suddenly happened on the traveling route during the evaluation time period.

In general, most of the vehicles traveled on congested signalized intersection network would experience all of stopped delay, approach delay, travel-time delay and time-in-queue delay. However, the vehicles traveled on non-congested signalized intersection network might not experience some part of stopped delay, approach delay, travel-time delay and time-in-queue delay occurred at the signalized intersections. Furthermore, the difference of travel times under non-congested traffic condition would be mainly caused by stopped delay occurred at the signalized intersection rather than the approach delay and/or travel-time delay. In this respect, the vehicles under non-congested traffic condition tend to maintain the higher speed than those under congested traffic condition so that the speed range of the vehicles traveled under non-congested traffic condition will be wider than that of the vehicles under congested traffic condition. Considering the characteristics of the speed data, the algorithm has the following procedures.

Step 1. Check the travel time range estimated by using the speed data collected during the same evaluation time period.

Step 2. Check the sum of red times of all traffic signals passed for a given specific approach.
If the range of travel times is smaller than the sum of red times, then the shortest one among the travel times measured during the evaluation time period can be used for proceeding to Step 4.
Otherwise, perform Step 3.

Step 3. If the range of travel times is greater than the sum of red times, then the longest one among the travel times measured during the evaluation time period can be used for proceeding to Step 4.

Step 4. Calculate speed value by using the selected travel time in Steps 2 and 3.

To validate the suggested algorithm, extensive data collection should be made. Currently, however, enough data were not collected for the validation of the proposed algorithm. Therefore, no performance test result is provided in this paper.

5 Concluding Remarks

It was confirmed that the travel speed values measured under non-congested traffic condition were severely fluctuated and the speed variation would be occurred by both the state of signals and the coordination of green times of signals. Thus, it is not reasonable to estimate the travel speed for representing the reality simply by using the average speed of all vehicles traveled during the same evaluation time period. The rationale of the algorithm proposed in this paper was developed based on the vehicles' trajectory in the time-space diagram. The basis of criterion for selecting the travel time to be used for final estimation of travel speed seems to be theoretically reasonable. The most useful advantage of the algorithm is that, the traffic analysts can reduce much tedious works for screening and selecting a bunch of speed data.

References

1. S. Lee, S. Oh and B. Son, "Heuristic algorithm for Traffic Condition Classification with Loop Detector Data," ICCSA 2005, LNCS 3481, pp.816-821, 2005
2. W.R. McShane and R.P. Roess, Traffic Engineering, Prentice-Hall, Inc. 1990
3. S. Lee and B. Son, "A Study of Measuring the Effectiveness for Road Safety Feature Improvement," The 4th Conference of Eastern Asia Society for Transportation Studies, 24-27 October, Hanoi, Vietnam, 2001
4. B. Son, H. Kim, C. Shin, and S. Lee, "Bus Arrival Time Prediction Method for ITS Application," KES 2004, LNAI 3215, pp.88-94, 2004

Home Network Observation System Using User's Activate Pattern and Multimedia Streaming

Kyung-Sang Sung[1], Dong Chun Lee[2], Hyun-Chul Kim[1], and Hae-Seok Oh[1]

[1]Dept. of Computer Science, KyungWon Univ., Korea
{pltofgod,dmzpolice78}@korea.com
[2] Dept. of Computer Science Univ., Korea
ldch@sunny.howon.ac.kr

Abstract. We suggest a simplified application model that could apply the information to the automated processing system after studying user's behavior patterns using authentication and access control for identity certification of users on ubiquitous computing environments. In addition, we compared captured video images in the fixed range by pixel unit through some time and checked disorder of them. According to experimental results, the proposed method is very reasonable to believe that we could extend various multimedia applications in our daily lives.

1 Introduction

Home networking technology is a new growth-power industry of the next generation that has interested people the most including DTV. It is the core technology field that will create enormous economic value for everyone by making all electric appliances of a household connected as one network unlike existing analogue electric appliances. Home network is a ubiquitous network environment in which a variety of electric appliances and many kinds of network exist all together [7].

Home gateway and home server is not only to offer high speed Internet service and multimedia service anytime by connecting the wired and wireless access network to the network of a household, but also to provide other functions like sharing resources within a household; all sorts of additional service such as entertainment, medical examination, and home shopping; operating the system by remote control with using a portable information unit; and home security service.

HAVi [3] based on IEEE1394 [6] for building a multimedia home network system and UPnP [4] middle-ware built to control and watch simple electric appliances are drawing people's attention now. In addition, a remote control observation system using digital video technology has enabled people to directly control the situation of the place far apart from the office. There are also more and more need to judge the situation quickly with the help of image or audio signs from the place far apart when there is an error in the controlling system, to check the condition of operation of each device, and to record image or audio data of unusual events in the form of text data and send them if necessary. In order to establish this kind of system, a remote control observation technique, an image processing technique, a distributed file storing technique, and a distribution controlling technique for preventing traffic are needed.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 74–80, 2005.

2 Related Works

Considering the ubiquitous computing environment [8], integrated management of ACL (Access Control List) by Home Gateway will be effective for controlling access in terms of stability. However, in case already certified users of devices require already fixed service, it is considered more effective to provide the right to access to the service by using the identification system within the device. Given this, we should discuss how to use and apply the technology of identification system considering the low level of performance of an information unit.

Embedded Linux technology [10] needs to be designed freely according to the user's demands and purposes, and the resources of the system need to be managed effectively. Embedded Linux that will be used for the system of home network connecting devices will include not only basic operating techniques such as anytime support, quick booting, electric power management, and multimedia file system support, but also embedded GUI, multimedia middle-ware and application skills, controlling middle-ware skills, VM technique and so on.

3 Proposed System

3.1 System Model

Fig. 1 shows the whole structure of the system. The server, which gets video signals through the module from a camera, sends the video data acquired from the server controller on the wired and wireless network after checking if there is a problem in the server state. Concerning the user who has demanded the data, the system extracts the user's monitored way of moving after identification from the user profile DB, and sends the data through the client module from DB manager. The video data is played in the appropriate form to the user's device.

The system will be used as a security solution to watch if there is a trespasser. This system sends image and audio data anytime with using RTP of JMF through an observation mode, and if the system finds a problem when comparing the image initially recorded and the one being sent anytime, it sends those two images through PDA or PC to the user. In the design of the server and client, the server gets data through the socket with the client, and the video signals are realized by using RTP of JMF. Real-time technique is very important in the controlling system for image streaming.

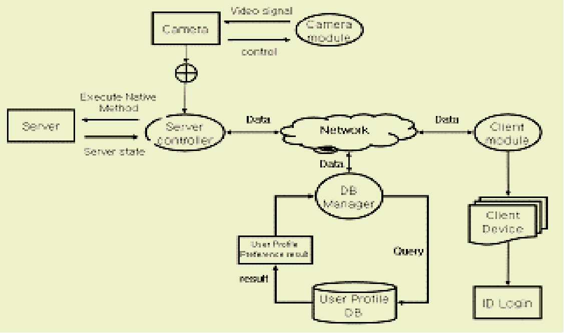


Fig. 1. The system structure

Therefore, we have also designed the process which can control and observe the system by using Java for making the platform independent in order to control the hardware.

The audio/video signals obtained through an observation mode are sent to the server, and the user is asked to send the audio/video signals through TCP/IP sockets after logging on to DB through web or PDA and passing the identification test [8, 10]. When the user online asks the system to send image data anytime, the system sends image data after changing the form of image files to PNG which the mobile device accepts after capturing images at a certain interval. When the system captures video signals at a certain interval and compares the image files at a unit of pixel, if there is an image that is beyond the range allowed, it is considered suspicious.

3.2 Algorism for Identifying Users

Message digest is a one-way conversion, so given the trait that it cannot be converted to the original data, it is used for confirming a password. Even if an attacker approaches the place where all passwords are stored to find out passwords, he cannot turn hashed passwords into readable signals. In order to prevent this kind of attack, the way to store a password with salt is being used these days. In that case, it is impossible for an attacker to find out the password by hashing it randomly. Fig. 2 shows the process of creating a password and confirming it with explanations.

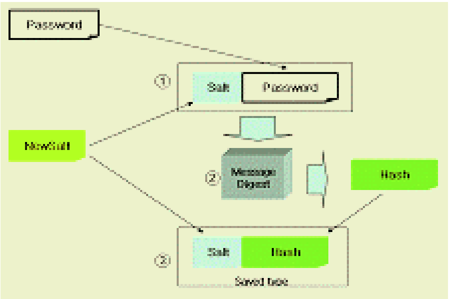
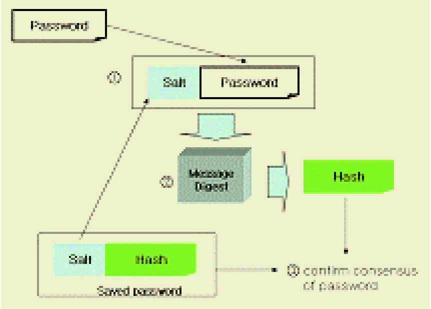
 Password ① Salt Password NewSalt ② Message Digest Hash ③ Salt Hash Saved table	 Password ① Salt Password ② Message Digest Hash Salt Hash Saved password ③ confirm consensus of password
① Create new random salt and put it in front of the password. ② Make it through the process of hashing. ③ Store salt and hashed data together in order to compare them when the password is input.	① Add salt of the stored password to the password that the user inputs. ② Hash salt and the password. ③ Compare the result of hashing and the data stored in file. Notice comparing the hashed result.

Fig. 2. The process of creating and confirming a password

It is to observe actions of a user comprehensively by using the monitor agent for the user's activate pattern analysis. In order to do this, the monitor agent monitors the user's actions comprehensively, and presents them in the abstract. The purpose of the monitor agent is to monitor, analyze, and use a user's activate pattern. Monitored data is divided into each user to make a profile of each user, and monitored data are put in the abstract as the user's activate pattern.

The purpose of an adaptive agent system is to understand the user's intention, to extract the information in which the user is interested, and to analyze the user's activate pattern accurately. The user of the home network system has to get the information of changed situations within the house as well as situations of the present.

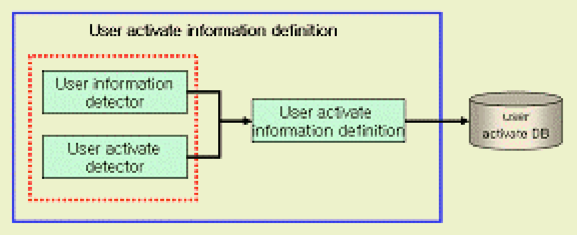

User activate information definition
User information detector
User activate detector
User activate information definition
User activate DB

Fig. 3. Monitor Agent for user's activate pattern analysis

The following algorithm shows the creating images anytime and objectifying the images.

```
// create of object about image
LoadImage BufferedImage(String filename){
          Create image("Create image object")
          mt.addImage(img, 0);
          try{
            mt.waitForID(0);
           }catch(InterruptedException e){
          System.exit(0); }
           if(mt.isErrorID(0)) System.exit(0);
            "img_Size(width, height)"
            Image_Convert();
            BufferedImage(width, height, " img_Type");
            "Image_Convert()"
            return }
      end LoadImage()
```

In the create_Img class, when the two images (of the present and the past) for comparing video signals are being created, images that uses time characteristics are created, and then the objects of the images are created. There should be an encoding process for the images in order to send the objects to an appropriate mobile device to the user. The following algorithm shows the checking the range of allowance.

```
// create of class for encoding process of the png
Image_Convert(current_Img, current_Time)
         {
          Img_con <-- Create_File.tag;
          try {
              String file <-- convert_Format(png);
              fos <-- new FileOutputStream(file);
              pe1 <-- new PngEncoderB(fImage);
              b <-- pe1.pngEncode();
              fos.write(b);
              fos.flush();
            fos.close();
            if(user.tag)  {
              ClientLoginForm.client.sendCurrentimage(file,currenttime);
            } else if(ClientLoginForm._mode == 3) {
              DirectPalm._palmhandler.sendCurrentimage(file,currenttime); }
   end Image_Convert()

// checks the range of allowance for the created two images
equal_Img(Real_Img_1, Real_Img_2)
    img ← false;
    case {
      Real_Img_1 = null and Real_Img_2 = null : img ← true;
      Real_Img_1 ≠ null and Real_Img_2 ≠ null : if Real_Img_1.tag = Real_Img_2.tag then {
              if (Real_Img_1.tag = 0) then img ← (Real_Img_1.data = Real_Img_2.data);
              else img ← equalList(Real_Img_1.data, Real_Img_2.data);
              if (img) then img ← equalList(Real_Img_1.link, Real_Img_2.link); } }
      eturn img;
end equal_Img()
```

In the Image_Convert class, the system encodes the images by using a pngEncode() function in an appropriate form to the user's mobile device. Then it checks the range of allowance for the images created at present and those that has been stored at a certain interval in the past. If the images go beyond the range of allowance, the system senses that there is a change in the situation of the house and sends an alarm signal to the user's device to notify the user of the present situation in the house.

4 Experimental Evaluation

The environment of development for the system is based on Pentium ▯ PC. It is also based on Window XP as an operating system and on Apache Web Server. It uses Oracle as a database. JDK 1.4 is used for its environment of development, and JMF engine is used for providing multimedia. J2ME is used for an embedded device that can be used when equipped to a mobile device. In addition, XML is used for providing information of each user and of each device. And the system has been tested through an emulator (i.e., PalmOS) in order to control information of the mobile device. The system gets a user's activate pattern analysis data from the server through IP and Port which has already been established after being connected to the server on a connection mode and passing an identification test.

The test environment in Fig. 4 is to show a situation of the inside of the house on an observation mode to the user by using a camcorder set up in the laboratory for testing the capacity of the system and by using the intelligent multimedia server. In addition, in order to solve the problems of resources and memories, the system selectively provides information of the Camcorder that the user has most recently used and most often used through personalized service, which has increased the efficiency of the system by 23%. Besides, using a microphone has enabled the system to obtain the information of the area that a camera cannot reach, which has widened the range of observation and improved the quality of observation. This system has achieved its goal to make people's life safer by observing situations of the inside of the house through PC or a mobile device by using the wireless transmission module by remote control, checking records of intruding the houses, sensing an emergency automatically, notifying the user of the emergency, and turning out to solve the problem.

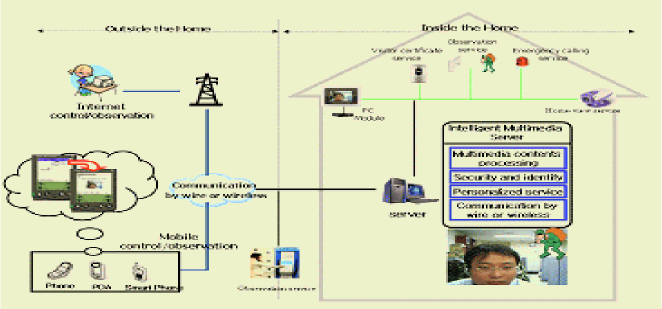

Outside the Home
Inside the Home
Visitor certificate service
Observation service
Emergency calling service
Internet control/observation
PC Module
Intelligent Multimedia Server
Multimedia contents processing
Security and identity
Personalized service
Communication by wire or wireless
server
Communication by wire or wireless
Mobile control/observation
Phone
PDA
Smart Phone

Fig. 4. Simulation environment

In table 1, as for the process of confirmation using symmetry-key encryption algorism for confirming between the server and the client, our self-confirmation technique using algorism of hashing the salt() function has proved that its speed is 30 times faster than P's technique based on PKI. Concerning stability of confirmation, however, P's system has turned out to be superior to the system we are suggesting. When normal confirmation is finished, both the user's activate pattern analysis data and the server's state data are sent to the user's device. Regarding realization of mobile devices, we have focused on connecting the data to the practical business by using the data renovated by the embedded system. Image data are sent to PDA at the average speed of 18 frames. When the user's pattern algorism is used, it saves resources and time by about 23%. Considering the extra memory of PDA, there is no problem in processing video signals that the server sends to PDA after comparing the signals.

Table 1. System Capacity Test (◎: well supported. △: quite supported, □: not supported)

Item	Proposed system	F's system	L's system	P's system
connection media	TCP/IP	PLC	PLC	TCP/IP
user's device	PDA, computer	mobile phone, computer	mobile phone, computer	mobile phone
confirmation process	self-confirmation	connecting server	connecting server	based on PKI
speed of confirmation process	5 ~ 10sec	18 ~ 30sec	10 ~ 15sec	30 sec↑
stability	△	□	□	◎
video data processing (frame)	18 Frame ↓↑	10 Frame ↓	15 Frame ↓	18 Frame↑
time for connecting after confirmation	10 sec↓	10 sec↓	8 sec↓	8 ~ 15sec↓
Resource (memory)	◎	□	□	□
personalized service	◎	□	□	□
audio data	◎	□	□	□

5 Conclusions

We suggested the remote controlling and anytime observation system for situations within the house based on TCP/IP and RTP. This controlling and observation system by remote control has a structure in which the client system that is connected to the place far apart controls and monitors situations of the place by remote control by using user's interface through his mobile device. This proposed system is expected to be effective in preventing industrial disasters in advance when applied to an industrial field.

References

1. PalmOS Emulator, http://www.palmos.com/dev/tech/
2. Sun Microsystems, "Jini Technology core Platform Specification," http://java.sun.com/d- eveloper/products/jini/index.jsp, jun.2003
3. HAVi Forum, "HAVi Specification 1.1," http://www.havi.org, May. 2001
4. UPnP Forum, "UPnP Device Architecture 1.0," http://www.upnp.org, Dec. 2003
5. http://www.ieee802.org
6. E. Callaway and et al., "Home Networking with IEEE802.15.4" IEEE Comm. Mag., 2002
7. Ken SAKAMURA, Ubiquitous Computing KAKUMEI, 2002
8. Jun Ho Park et al, "Multi-Agent System absed Home Network Management System using Extended Tuple Space Model," SAM'02 SCOPE
9. Tim Kindberg, Armondo Fox, "System Software for Ubiquitous Computing," Pervasive Computing, 2002
10. Yu Feng, Dr.Jun Zhu, Wireless Java programming with J2ME, Sams, 2001

Determination of Optimal Locations for the Variable Message Signs by the Genetic Algorithm

Jaimu Won[1], Sooil Lee[1], and Soobeom Lee[2]

[1] Dept. of Urban Engineering, Hanyang Univ., Korea
drwon21@hanmail.net, sooillee@korea.com
[2] Dept. of Transportation Engineering, Univ. of Seoul, Korea
mendota@uos.ac.kr

Abstract. The Variable Message Signs (VMS) are useful way to reduce the socio-economic costs due to the traffic congestions and delays by providing the information on traffic condition to drivers. This paper provided a methodology to determine the locations of VMS's in terms of the minimization of the delay by applying the genetic algorithm. The optimal number of VMS's was also derived by the economic analysis based on the cost and the benefit. The simulation considered the variation of traffic volume, the frequency and duration of the incident, and the traffic conversion in order to reflect the real situation.

1 Introduction

The socio-economic cost by traffic congestion has been increased seriously for past decades. In order to lessen the problem, extensive efforts have been given to make the operation of the prevailing facilities more effective. Among these, the variable message signs have contributed a lot by reducing the travel times. The VMS's can provide the information on the delays or the incidents on the down stream, and the road users can select an alternative route.

This paper aimed to determine the optimal locations of VMS's in terms of the maximization of the socio-economic benefit by applying the genetic algorithm which is a part of the artificial intelligence technique.

2 Literature Review

The previous studies on the VMS can be classified into two fields: the effect and the strategy of the information provision by VMS and the maximization of effect by the implementation of VMS.

Wardman, Bonsall and Shires evaluated the results of the route selection according to the VMS information by the state of preference method (1998). The drivers selected the detouring routes by the satisfaction rate of the VMS message, the cause and the duration of delay, the characteristics of drivers, and the familiarity for the highway network.

Seok and Choi studied for the location of VMS in order to establish the strategy for the provision of traffic information corresponding to the increase of the user's demand. They determined the priority of the routes by the analytic hierarchy process with the survey results from the traffic experts. In addition, the driver's behaviors were analyzed with a internet-based simulator composed of eight kinds of scenarios.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 81–86, 2005.

3 Approach of Research

The genetic algorithm (GA) applied in this study is a search algorithm based on the genetic mechanism of the biological laws of the survival of the fittest or natural selection. The GA is a useful methodology for the optimization. It does not provide a single optimal solution but the multiple ones by the priority.

The GA was applied to select the optimal locations of VMS's with the object function minimizing the highway user's delays in this study.

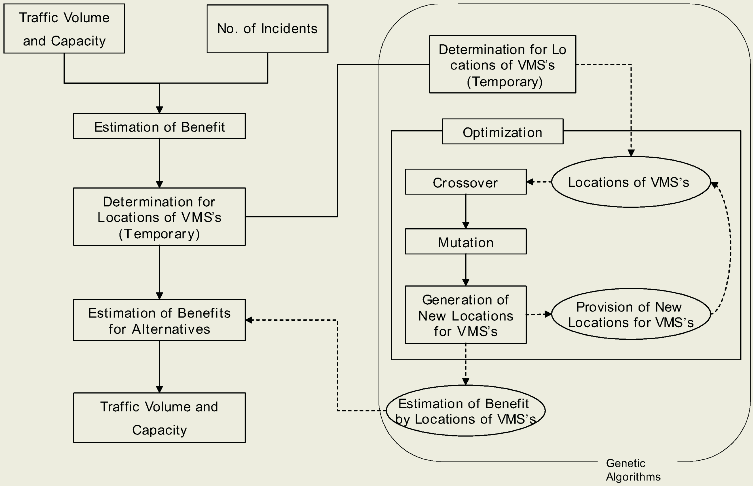

Traffic Volume and Capacity
No. of Incidents
Estimation of Benefit
Determination for Locations of VMS's (Temporary)
Estimation of Benefits for Alternatives
Traffic Volume and Capacity
Determination for Lo cations of VMS's (Temporary)
Optimization
Crossover
Locations of VMS's
Mutation
Generation of New Locations for VMS's
Provision of New Locations for VMS's
Estimation of Benefit by Locations of VMS's
Genetic Algorithms

Fig. 1. Procedure for the Determination of Optimal Locations for VMS's

3.1 Determination of Optimal Locations for VMS's

The benefit resulted from the implementation of VMS s shown in Figure 2. The incident occurred at A caused the delay amounted to the area of ΔABC. However, when the VMS is implemented, the delay decreases to ΔDEC. The size of ΔDEC depends on the location of the VMS.

The potential benefit was estimated with the variation of traffic volumes and the number and durations of the incidents, which is a crucial factor for the determination of optimal location for VMS. The potential benefit means the reduction of travel cost, and estimated with the BPR (Bureau of Public Road) function as following.

$$t_a = t_a^0 \left(1 + \alpha \left(\frac{V_a}{C_a} \right)^{\beta} \right)$$

t_a : travel time on a link a with a give traffic volume,

t_a^0 : travel time on a link a with no traffic,

V_a : traffic volume on link a,

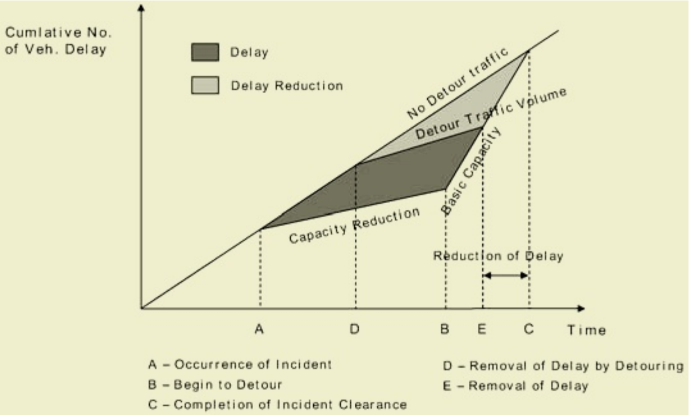


Fig. 2. Delays by Incident

C_a : capacity of link a

α, β : parameters (α=0.93, β=1.80)

The following friction function was also applied to determine the total travel cost on the route.

$$tt_a = t_a \times V_a$$

tt_a : total travel time for the total traffic volume on link a

3.2 Optimal Number of VMS's

The optimal number of VMS was determined by the economic analysis such as the benefit-cost ratio and the net present value based on the implementation cost and the socio-economic benefit (reduction of delays). The analysis period was ten years and the discount rate was 7.5%. The time value was employed from the research carried out at the Korea Development Institute.

Table 1. Time Value for Business Travel

Classification	Business Type		
	Auto Driver	Bus Driver	Truck Driver
Average Income (won/month)	1,427,000	1,194,832	1,090,531
Working Hours (hr/month)	119.2	200.3	
Wage per Hour (won/man-hr)	7,164	5,965	5,444
Overhead Rate for Wage (%)	29.9	26.7	36.3
Time Value (won/man-hr)	9,306	7,558	7,420

Source: Korea Development Institute, "Manual of Preliminary Economic Feasibility Study of Highway Investment Project (Revision)", 2000

4 Case Study

The subject site for the case study was a part of the National Highway 1. The traffic situations were simulated considering the variation of traffic volume, traffic conversion, and the type of incidents.

4.1 Subject Site

The subject site was a part of the National Highway 1 of four to eight lanes and 35km in length.

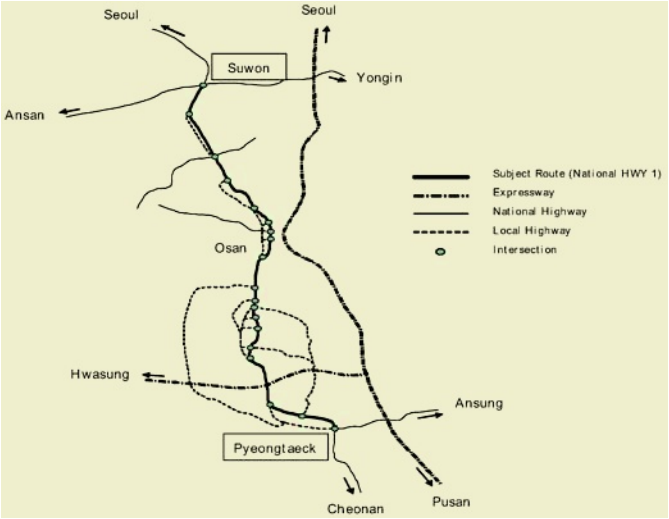


Fig. 3. Subject Site

4.2 Traffic Volume

In the computer simulation, three situations by traffic volume were prepared. The case 1 was for the actual prevailing traffic volume, and cases 2 and 3 are for the traffic volumes of 20% less and 20% more than prevailing conditions, respectively.

4.3 Traffic Volume Conversion

The traffic volume conversion could be defined as the detouring traffic volume which is derived with the probabilities of the driver's information perception and the decision for detour, as shown below.

Traffic Volume Conversion (DT) = NV × SP × DP

NV : link volume on the subject site,
SP : probability for drivers to percept the VMS information,
DP :probability of driver's decision to detour.

In this study, the rate of traffic volume conversion was 34.9% which was provided by a previous study.

Table 2. Cases for Simulation

Classification	Total Traffic Volume on the Network (Veh/Hr)	Location of Incident	Duration of Delay by Incident
Case 1	8,762	Random Generation among 7 Spots	Selection of Capacity Reduced to 10%, 30% or 50% Randomly
Case 2	7,952	Random Generation among 7 Spots	Selection of Capacity Reduced to 10%, 30% or 50% Randomly
Case 3	9,572	Random Generation among 7 Spots	Selection of Capacity Reduced to 10%, 30% or 50% Randomly

4.4 Number of Incidents and Delays

The number of incidents was estimated in the form of the probability due to the uncertainty of location and the frequency. The delays by the incidents were calculated corresponding to the random reduction of capacity to 10%, 30%, or 50%. The locations of incidents were selected one of seven spots pre-designated based on the accident record data.

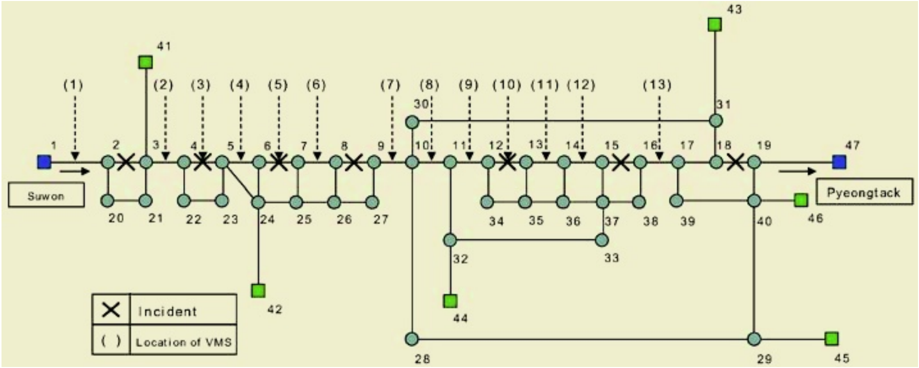


Fig. 4. Locations for Incidents and Probable VMS Implementation

4.5 Locations and Numbers of VMS's

The optimal locations of VMS's from the genetic algorithm and the optimal number by the economic analysis were provided by each case. The output provided for one to 13 VMS's by each case, and the final result is shown in Table 3. The number of VMS's on the result is the minimum with which the B/C ratio was greater than 1.0. It means that the effectiveness may be maximized by implementing additional VMS's when the prevailing number is less than that in the result.

Table 3. Optimal Number of VMS's by Case

Case	Optimal Locations of VMS's	Benefit by VMS (million won)	VMS Cost (million won)	Economic Feasibility	
				B/C	NPV
Case 1	12,2,4,1,7,9	11,865	1,560	2.0	253
Case 2	12,7,1,9,2,13	9,785	1,560	1.1	20
Case 3	12,1,7,9,2,13,6	17,302	1,820	1.4	103

5 Conclusions

This paper developed a methodology to determine the probable location of VMS's applying the genetic algorithm. The optimal locations of VMS's could be determined by the economic feasibility analysis with the cost and the potential benefit. The computer simulation was performed considering the variation of traffic volume, traffic volume conversion and the frequency and the duration of the incidents in order to reflect the actual situation. The methodology suggested in this paper may work as a useful tool for the determination of location for VMS implementation.

This paper assumed that the behaviors of drivers were homogeneous and constant, and focused on the locations and number of VMS's only. It is then desirable to extend the study on the probabilities for the perception of traffic information and route selection.

References

1. Kang, H. I., and C. H. Park, "Development of Classified Congestion Functions", Journal of Korean Society of Transportation Vol.16, No.2, 1998
2. Seo, C. Y., "Fundamental Study for Drivers ' Route Preference with Traveler Information", Journal of the Korea Planners Association Vol.30, No.3, 1995
3. Mammar, S., A. Messmer, P. Jensen, H. Haj-Salem, M. Papageorgiou and L. Jensen, "Automatic Control of Variable Message Signs in AALBORG", Transportation Research, Vol. 4, No. 3, 1996.
4. Wardman, M., P.W. Bonsall and J.D. Shires, "Driver Response to Variable Message Signs: A Stated Preference Investigation", Transportation Research, Vol. 5, No. 6, 1998.
5. Rama, P., and R. Kulmala, "Effects of Variable Message Signs for Slippery Road Conditions on Driving Speed and Headways," Transportation Research, 2000.
6. Luoma, J., P. Rama, M. Penttinen, V. Anttila, "Effects of Variable Message Signs for Slippery Road Conditions on Reported Driver Behaviour", Transportation Research, 2000.
7. Chatterjee, K., N.B. Hounsell, P.E. Frimin, P.W. Bonsall, "Driver Response to Variable Message Sign Information in London", Transportation Research, 2002.

Estimation of the Optimal Number of Cluster-Heads in Sensor Network

Hyunsoo Kim[1], Seong W. Kim[2], Soobeom Lee[3], and Bongsoo Son[4]

[1] Postdoctoral Researcher, Dept. of Mathematics,
Sungkyunkwan Univ., Korea
bayes1@hanmail.net

[2] Assistant Professor, Div. of Applied Mathematics,
Hanyang Univ., Korea
seong@hanyang.ac.kr

[3] Assistant Professor, Dept. of Transportation Engineering,
Univ. of Seoul, Korea
mendota@uos.ac.kr

[4] Associate Professor, Dept. of Urban Planning and Engineering,
Yonsei Univ., Korea
sbs@yonsei.ac.kr

Abstract. A sensor network system consisting of a large number of small sensors with low-power can be an effective tool for collection and integration of data by each sensor in a variety of environments. The collected data by each sensor node is communicated through the network to a single base station that uses all collected data to determine properties of the data. Clustering sensors into groups, yields that sensors communicate information only to cluster heads and then the cluster-heads communicate the aggregated information to the base station. We estimate the optimal number of cluster-heads among randomized sensors in a bounded region. We derive solutions for the values of parameters of our algorithm that minimize the total energy spent in the wireless sensor network when all sensors communicate data from the cluster-heads to the base station. Computer simulation shows that the energy consumption reduce as the optimal number of cluster-heads for the proposed method.

1 Introduction

Recent developments in wireless sensor network have motivated the growth of extremely small and low-cost sensors that possess sensing, signal processing and wireless communication capabilities. These sensors can be expended at a cost much lower than conventional wired sensor systems. Each sensor is capable of detecting conditions around areas of distributed sensors such as temperature, sound, or the presence of certain objects. Sensor network system has gained increasing importance due to their potential benefits for some civil and military applications such as combat field surveillance, security and disaster management. The smart dust project at university of California, Berkeley [6, 8, 12] and WINS project at UCLA [9] are attempting to build such extremely small sensors that

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 87–94, 2005.

entitle autonomous sensing and communication in a cubic millimeter. These systems process data gathered from multiple sensors to monitor events in an area of interest.

[2] have analyzed the capacity of wireless ad-hoc networks and derived the critical power at which a node in a wireless ad-hoc network should communicate to form a connected network with probability one (cf. [3]). A sensor in wireless sensor network can communicate directly only with other sensors that are within a radio range in a cluster. To enable communication between sensors not within each communication range, the sensors form a new cluster in distributed sensors.

Sensors in these cluster detect events and then communicate the collected information to a cluster-head. The cost of transmitting a bit is lower than other routing protocol methods (minimum transmission energy routing protocol, direct communication protocol, and *etc*). Prolonged network lifetime, scalability, and load balancing are important requirements for many wireless sensor network applications.

In the minimum energy routing protocol, sensors route data are destined ultimately for the base station through intermediate sensors. The problem of these protocol is only to consider the energy of the transmitter and neglect the energy consumption of the receivers in determining the route. As distances between nodes change, energy consumption can be reduced. The low-energy adaptive clustering hierarchy (LEACH) includes the use of energy-conserving hardware. Moreover, a higher lifetime of sensor networks can be accomplished through optimized applications, operating systems, and communication protocols.

We concentrate on the number of cluster-heads among all distributed sensors in an interested region as distances between sensors. [10] cited an expected distance between a cluster-head and sensors [1]. However we consider an expected distance between a cluster-head and sensors with a radio range of the cluster-head in a cluster and an expected distance between cluster-head and base station outside a wireless sensor network system. The network system can determine, a priori, the optimal number of cluster-heads to obtain in a region of distributed sensors. And we consider the distribution of the energy consumption of sensors. The essential operation in sensor clustering is to select a set of cluster-heads among the sensors in the network, and cluster the rest of the sensors with these cluster-heads. cluster-heads are responsible for coordination among the sensors within their clusters, and communication with each non-cluster-heads in clusters.

The rest of the paper is presented as follows: In Section 2, we review the concept of LEACH. In Section 3, we propose the estimation of the number of cluster-heads with hyper-parameters in the Poisson process. In Section 4, we demonstrate effectiveness of the proposed optimal number k of cluster-heads by computer simulation. We provide a conclusion in Section 5.

2 Preliminaries

2.1 LEACH Protocol Architecture in Wireless Sensor Network

LEACH is a self organizing, adaptive clustering protocol that uses randomization to distribute the energy consumption evenly among the sensors in the network. In LEACH, the sensors organize themselves into local clusters, with one node acting as the local base station or cluster-head. If the cluster-heads were chosen a priori and fixed throughout the system lifetime, as in conventional clustering algorithms, it is know that the selected cluster-heads would die quickly, ending the useful lifetime of all nodes belonging to those clusters. LEACH includes randomized rotation of the high energy cluster-head such that it rotates among all sensors in order to not spent the energy of a specific sensor. In addition, LEACH carry out local data fusion to compress the amount of data being sent from cluster-heads to the base station, moreover, reducing energy consumption and increasing sensor lifetime. Then LEACH prolong the lifetime of the network system. Clusters can be formed based on many properties such as communication range, number and type of sensors and geographical location. Figure 1 depicted data communication between sensors and the base station.

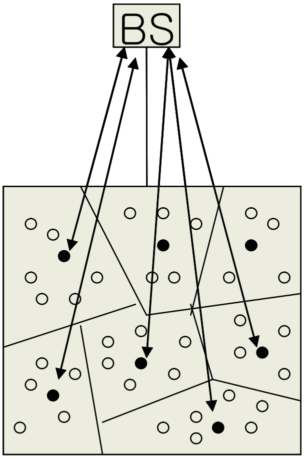


Fig. 1. Communication with cluster-heads and base station

2.2 Clustering Algorithm

Cluster-heads in the sensor network advertises itself as a cluster-head to the sensors within its radio range. This advertisement is forwarded to all the sensors that are no more than radio range away from the cluster-head. Any sensor that receives such advertisements and is not itself a cluster-head joins the cluster of the closest cluster-head. Any sensor that is neither a cluster-head nor has joined any cluster itself becomes a cluster-head; we call these cluster-heads the forced cluster-heads. Because we have limited the advertisement forwarding to radio range, if a sensor does not receive a cluster-head advertisement within time duration t (where t units is the time required for data to reach the cluster-head from any sensor radio range away) it can infer that it is not within radio range of any volunteer cluster-head and hence become a forced cluster-head. Moreover, since all the sensors within a cluster are at most radio range away from the cluster-head, the cluster-head can transmit the integrated information to the base station after every t units of time. This limit on radio range allows the cluster-heads to schedule their transmissions. The energy consumption used in the network for the information gathered by the sensors to reach the base station will depend on the number of cluster-heads and radio range r of our algorithm. Because of organizing the sensors in clusters to minimization of energy consumption, we need to find the number of cluster-heads in our algorithm that would ensure minimization of energy consumption. The basic idea of the derivation of the optimal number of cluster-heads is to define a function for the energy consumption used in the network to communicate information to the base station and then find the number of cluster-heads minimizing it. We derive the number of cluster-heads in a given environment. We need computations of the optimal number of cluster-heads in a cluster.

2.3 Radio Energy Consumption Model

We assume a simple model for the radio hardware energy consumption where the transmitter consumes energy to run the radio electronics and the power amplifier, and the receiver consumes energy to run the radio electronics. For the experiments described here, both the free space (d^2 power loss) and the multipath fading (d^4 power loss) channel models were used, depending on the distance between the transmitter and receiver. Power control can be used to invert this loss by appropriately setting the power amplifier-if the distance is less than a threshold d_0, the free space (fs) model is used; otherwise, the multipath (mp) model is used. Thus to transmit an l-bit message a distance d, the radio expends

$$\begin{aligned} E_T(l,d) &= E_{T-elec}(l) + E_{T-amp}(l,d) \\ &= \begin{cases} l * E_{elec} + l\epsilon_{fs}d^2, d < d_0 \\ lE_{elec} + l\epsilon_{mp}d^4, d_0 \leq d' \end{cases} \end{aligned}$$

and to receive this message, the radio expends:

$$E_R(l) = E_{R-elec}(l) = lE_{elec}.$$

The electronics energy, E_{elec}, depends on factors such as the digital coding, modulation, filtering, and spreading of the signal, whereas the amplifier energy, $\epsilon_{fs}d^2$ or $\epsilon_{mp}d^4$, depends on the distance to the receiver and the acceptable bit-error rate.

3 Estimation of the Optimal Number of Cluster-Heads

In LEACH, the cluster formation algorithm was created to ensure that the expected number of clusters per round is k, a system parameter. We can analytically determine the optimal value of k in LEACH using the computation and communication energy models.

Let R denote a bounded region as a square of a side $2a$. Let $X(R)$ be the number of sensors contained in a region R. Assume $X(R)$ is distributed uniformly in R. Let d_{toBS} be a random variable that denotes the length of the segment from a sensor in R. We assume that the base station is located near the sensor network system. Then the expected distance from base station to sensors is given by

$$E[d_{toBS}] = \int\int_R \sqrt{x^2 + (y - y^*)^2}\frac{1}{4a^2}dxdy. \tag{1}$$

The expected distance from the base station to sensors depend on the parameter y_1 where y_1 is the position of the base station outside a network system.

Suppose that we have k clusters, there are on average N/k nodes per cluster (one cluster-head and $(N/k) - 1$ non- cluster-head nodes). Each cluster-head consumes energy receiving signals from sensors, aggregating the signals, and transmitting the aggregate signal to the base station. Since the base station is far from sensors, presumably the energy consumption follows the multi-path model. Therefore, the consumed energy in the cluster-head per a bit of data during a single frame is

$$E_{CH} = l \times [(\frac{N}{k} - 1)E_{elec} + \frac{N}{k}E_{DA} + E_{elec} + \epsilon_{mp}(d^*_{toBS})^4],$$

where l is the number of bits in each data message and we have assumed perfect data aggregation and $d^*_{toBS} = E[d_{toBS}]$.

Let $X(R_{CH})$ be the random variable denoting the number of sensors except a cluster-head in a cluster and R_{CH} be a square with a side $2a/\sqrt{k}$. Let d_{toCH} be the distance of segment connecting the sensor to the cluster-head in a cluster. Assume the cluster-head located in the center of a cluster. Then according to results in [1], the expected number of non-cluster-heads and the expected length from the sensors to the cluster-head in a cluster are given by

$$E[X(R_{CH})|X(R) = N] = \frac{N}{k} - 1 \tag{2}$$

and

$$\begin{aligned} E[d_{toCH}|X(R_{CH}) = N/k] &= \int\int_{R_{CH}} \sqrt{x^2 + y^2}k(x,y)dxdy \\ &= \frac{0.7652a}{k^{1/2}}, \end{aligned} \tag{3}$$

respectively, where the density $k(x,y)$ of sensors follows a uniform distribution in the occupied area by each cluster that is approximately $4a^2/k$.

Each non-cluster-head node only needs to transmit its data to the cluster-head once during a frame. Presumably the distance to the cluster-head is small, so the energy consumption follows the Friss free-space model (d^2 power loss). Thus, the energy used in each non-cluster-head mode is

$$E_{non-CH} = l \times [E_{elec} + \epsilon_{fs}(d^*_{toCH})^2],$$

where $d^*_{toCH} = E[d_{toCH}|X(R_{CH}) = N/k]$.

The energy dissipated in a cluster during the frame is

$$E_{R_{CH}} = E_{CH} + (\frac{N}{k} - 1)E_{non-CH},$$

and the total energy for the system is

$$\begin{aligned} E_{Total} &= k \times E_{R_{CH}} \\ &= l \times [(2N-k)E_{elec} + NE_{DA} + k\epsilon_{mp}(d^*_{toBS})^4 + (N-k)\epsilon_{fs}(d^*_{toCH})^2]. \end{aligned} \quad (4)$$

Here, E_{Total} is minimized by a value of k that is a solution of the first derivative of (4). The second derivative of (4) is positive and log concave for the only real root in the first derivative of (4) and hence it minimizes the total energy consumption.

We can find the optimal number of clusters as following:

$$k_{opt} = \left\{ \frac{0.5855N\epsilon_{fs}a^2}{\epsilon_{mp}(d^*_{toBS})^4 - E_{elec}} \right\}^{1/2}. \quad (5)$$

4 A Simulation Study

We conduct a simulation with the algorithm described in Section 2 based on networks of sensors distributed uniformly. We used a N- sensors network where sensors were randomly distributed between $\{x|-50 < x < 50\}$ and $\{y|-50 < y < 50\}$ with the base station at location $(x = 0, y = y^*)$. The communication energy parameters are set as: $E_{elec} = 50$ nJ/bit, $\epsilon_{fs} = 10$ pJ/bit/m^2, $\epsilon_{mp} = 0.0013$ pJ/bit/m^4, and $a = 50$. The energy for data aggregation is set as $E_{DA} = 5$ nJ/bit/signal. See [11] Let y^* be a location between $75m$ and 185 [5]. Then $57 < d^*_{toBS} < 163$ by (1). When $N = 100$, the optimal number of cluster-heads is between 1 and 11 by (5). When $N = 200$, the optimal number of cluster-heads is between 1 and 15 by (5). We show that the optimal number of cluster-heads determines as d^*_{toBS}:when $N = 100$, $(k_{opt}, d^*_{toBS}) = (3, 105), (5, 81), (6, 74), (10, 57)$. Figure 2 and 3 are depicted the optimal number of cluster-heads minimizing the total energy consumption as the number of sensors and the distance between base station and sensor network system. Thus we know the number of cluster-heads depend on the distance between base station and sensor network system.

In this network system the expected distance between a sensor and the base station is only dependable on a given bounded region. However, the expected distance between a sensor and a cluster-head in a cluster is dependable on the

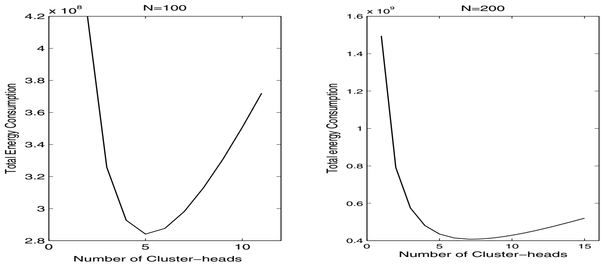


Fig. 2. Total energy consumption with $d^*_{toBS} = 81$

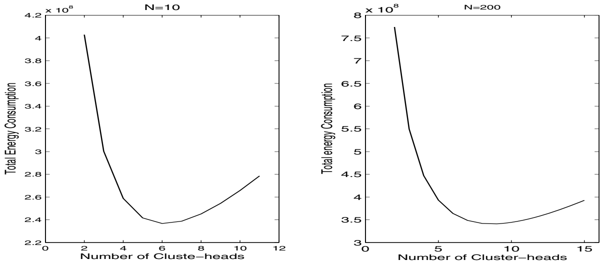


Fig. 3. Total energy consumption with $d^*_{toBS} = 74$

number of sensors and the number of clusters in a bounded region. Thus the expected distance between a sensor and a cluster-head in a cluster decreases as increasing of number of sensors in a bounded region. Therefore, this can be attributed to the fact that the optimal number of cluster-heads determines as the different density of sensors in a bounded area and total energy consumption minimizes as number of cluster-heads in the sensor network.

5 Conclusion

The sensors which become the cluster-head in LEACH architecture spend relatively more energy than other sensors because they have receive information from all the sensors within their cluster, aggregate this information and then communicate to the base station. Hence, they run out of their energy faster than other sensors. We have found the optimal number of cluster-heads for the proposed algorithm that minimize the energy spent in the network, when sensors are uniformly distributed in a bounded region. We know the number of cluster-heads depend on the distance between base station and sensor network system.

Acknowledgements

This work was supported by grant No. R01-2005-000-10141-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

References

1. S. G. Foss and S. A. Zuyev, “On a certain segment process with Voronoi clustering,” INRIA, Rapport de Recherche No, 1993.
2. P. Gupta and P. R. Kumar, “The capacity of wireless networks,” IEEE Transaction on Information Theory, Vol. IT-46, No. 2, 388-404, March, 2000.
3. P. Gupta and P. R. Kumar, “Critical power for asymptotic connectivity in wireless networks,” Stochastic Analysis, Control, Optimization and Applications: A Volume in Honor of W. H. Fleming, 547-566, 1998.
4. W. Heinzelman et. al, “Energy-efficient communication protocol for wireless sensor networks,” in the Proceedings of the 33rd Hawaii International Conference on System Sciences, Hawaii, 2000.
5. W. Heinzelman et. al, “An application-specific protocol architecture for wireless microsensor networks,” IEEE Transaction on Wireless Communications, Vol. 1, No. 4, Oct. 2002.
6. V. Hsu, J. M. Kahn, and K. S. J. Pister, “Wireless Communications for Smart Dust”, Electronics Reaserch Laboratory Technical Memorandum M98/2, Feb. 1998.
7. C. Intanagonwiwat et. al, “Directed diffusion: A scalable and robust communication paradigm for sensor networks,” in the Proceedings of MobiCom’00, MA, 2000.
8. J. M. Kahn, R. H. Katz and K. S.J. Pister, “Next Century Challenges: Mobile Networking for Samrt Dust,” in the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom 99), 271-278, Aug. 1999.
9. G. J. Pottie and W. J. Kaiser, “ Wireless integrated network sensors,” Communications of the ACM, Vol 43, No. 5, 51-58, May, 2000.
10. Seema Bandyopadhyay and Edward J. Coyle, “An energy efficient hierarchical clustering alogoritthm for wireless sensor networks,” IEEE INFOCOM 2003 - The Conference on Computer Communications, vol. 22, no. 1, 1713-1723, Mar 2003.
11. A. Wang, W. Heinzelman, and A. Chandrakasan, “Energy-scalable protocols for battery-operated microsensor networks,” Proc. 1999IEEE workshop Signal Processing Systems, 483-492, Oct. 1999.
12. B. Warneke, M. Last, B. Liebowitz, Kristofer and S. J. Pister, “Smart Dust: Communication with a Cubic-Millimeter Computer,” Computer Magazine, Vol. 34, No 1, 44-51, Jan. 2001.

Development of Integrated Transit-Fare Card System in the Seoul Metropolitan Area

Jeonghyun Kim[1] and Seungpil Kang[2]

[1] Research Prof., Dept. of Civil & Environmental Eng., Hanyang Univ., Korea
jeonghkim@hanyang.ac.kr
[2] Prof., Dept. of Urban Engineering, Seoul National Univ., Seoul, Korea
kangsp@gong.snu.ac.kr

Abstract. Urban traffic congestion is one of the most serious problems in many large cities of the world. The traffic demand management is the most practicable countermeasure, and the role of urban transit is increasing. The electronic fare card system can enhance the service quality of transit in terms of users' convenience and the speed of operation. In addition, the importance of the inter-modal transit service is increasing, and the fare system between two different modes of transportation is a crucial issue. The city of Seoul in South Korea has recently introduced the integrated transit-fare card system, corresponding to the reform of the urban bus system. This paper focuses on the basic architecture of the fare card system.

1 Introduction

Seoul, South Korea is one of the largest cities in the world, with a population of 12 million, including the surrounding suburbs. Public transportation in the central city of Seoul alone is estimated at 29.6 million trips per day. Thirteen subway lines provide service for daily trips reaching over 8.2 million. In addition, private buses served more than 400 routes until June 2004. From July 2004, the city of Seoul introduced a "semi-public bus operating system" that retains private bus firms, but leaves routes, schedules, and fare decisions to the Seoul Metropolitan Government. In allowance with the reform, the city introduced the integrated transit-fare card system (the T-Money System), which can be used not only for buses, but also for the subway. Moreover, the system credits fare exemption or a discount when transferring between multiple modes of transportation. The collected fare is distributed to the bus and subway service providers according to the portion of riders. The specifications have been defined, and the system is standardized.

Hong-Kong, Singapore, and Taiwan have been introduced to similar systems. But the fare exemption or the discount when transferring is not accepted except in Taiwan, and the Easy Card of Taiwan has not been standardized.

The T-Money system may be the most advanced and best case towards the implementation of the integrated transit-fare card system. Introducing the basic architecture of the integrated transit-fare system of Seoul, this paper identifies the tasks for future enhancement of this system.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 95–100, 2005.

2 Structure of the T-Money System

The T-Money system is comprised of the fare adjustment office, the network, an RF reader, and the T-Money card, or other affiliated cards such as credit cards, prepaid cards and debit chips in mobile telephones. The network links the card (RF transponder), RF reader, and the adjustment office via the wire and/or wireless systems. Figure 1 displays the overview of the T-Money system. The system has been designed with an open architecture, then the specifications are released.

The T-Money cards contain the CPU. The memory capacity is 120Kb and the B/L capacity is 600 million (64Mb). The cards are designed in accordance to the EMV standard.

3 Data Transmission System

When the cards come into contact with the RF readers, the RF readers calculate the fare in accordance with the previous boarding/disembarking records stored on the cards. For credit cards, the fares are reported to the card company on the same day. In the case of prepaid cards, the fares are subtracted from the pre-charged amount. Following these processes, the records stored on the cards and the status records of the RF readers generate the event (boarding/ disembarking) records, and the event records are then stored on the RF reader. Simultaneously, the RF readers record the collected amount and the status of RF readers, such as the time, location, and etc.

The event records are stored on an RF file, and transmitted periodically to the T-Money adjustment office. Transmission for subways occurs once every 3,000 events or every 5 minutes, and the event records on buses are collected every time the buses return to their starting depots.

The event records collected on the bus disembarking RF reader are transmitted to the boarding RF reader, and both boarding/disembarking records are stored by the operation unit. When the buses return to the depot, the stored records are moved to the main computer of the bus companies, and the records are transmitted to the T-Money adjustment office through an on-line network.

For the subway system, the event records collected on the computer of each station is in real time, and the files are transmitted to the adjustment office via servers of each subway company after every 3,000 events or 5 minutes.

4 Contents of Event

The event records are composed of three types of information: the present boarding/ disembarking records, the previous boarding/ disembarking records, and the operation records. The boarding/ disembarking records include the route information, the stations and times of boarding/disembarking, card numbers, passenger type (ie. student/senior citizen discounts), transfer records, and the fares. The previous boarding/disembarking records consist of the route information, vehicle number, boarding/ disembarking stations and times, and the number of transfers. The bus and subway operation records are composed of driver information, departure time by station, numbers of passenger boarding and disembarking (in the case of disembarking tags only) by date, route, vehicle, operation and passenger type.

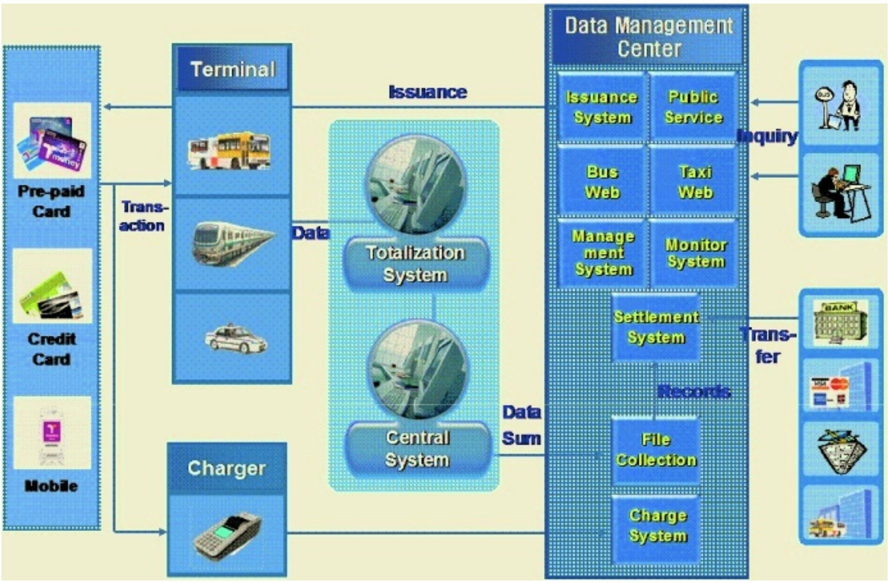


Fig. 1. Overview of the T-Money System

5 Data Processing System

5.1 Data Processing Unit

There are three separate data processing units, which are the event, record, and transaction units. The event unit processes a record upon boarding/disembarking. It records the number of the T-Money fare card, the station and time of boarding/disembarking, the fare amount, and the number of transfers (transactions). These events are transmitted to the adjustment office where the boarding and disembarking records are paired up. The record unit documents the boarding/disembarking information. And transferring the records generated from the first boarding to the final transfer, involves the transaction unit.

5.2 Data Processing

An information unit (event) is the information related to boarding and disembarking generated at the moment the card is tagged. The information generates files which are transmitted to the adjustment office through an on-line network. The boarding/ disembarking records are prepared by coupling the origins and destinations, then stored on the database.

For new events of boarding/disembarking, the corresponding disembarking/boarding event is checked on whether it is already on the database. If the event is stored on the database already, the boarding/ disembarking record is merged and updated. In the case that the event is new, a new record is inputted to the database.

During the adjustment process, the records on the database are grouped by the transaction unit, and the adjustment is performed for the finalized transaction units.

6 Derivative Traffic Information

6.1 Generation of Traffic Information

The unit of information collected and stored by T-Money is a record of boarding and disembarking. This single fare collection system is required in order to implement the discount system during transactions. The individual records of consecutive boarding/disembarking are necessary in order to obtain the required records for distribution of the income to the various and multiple transit operators.

The record of an individual's consecutive boarding/ disembarking makes it possible to verify the individual transit user's movement by time period in a day. This information can be used as the basic data in the development of an origin-destination (O-D) matrix by period in a day around the Seoul Metropolitan area. In addition, the movement pattern of residents in the area can also be determined. The O-D matrix is the most important data source for transportation planners.

A great number of residents and the economically active population in Seoul use the transit linked with the T-Money system. It is a useful tool in defining the movement pattern by time period in a day, and it becomes an incredible asset in urban planning, environmental policy, and economic policy. Not only does this benefit the city of Seoul and the local government surrounding the city, but it also benefits the central government.

6.2 Types of Traffic Information

The types of traffic information available can be classified into five categories as shown in Table 1.

Table 1. Types of Traffic Information Derived from the T-Money System

Type of Traffic Information	Contents
Location	• Movements made by station, month, day of week, time period in a day • Movements made by location, month, day of week, time period in a day • Movements made by region, month, day of week, time period in a day
Inter-Regional	• Movements and travel time between stations by month, day of week, time period in a day • Movements made by location, month, day of week, time period of a day • Movements made by region, month, day of week, time period of a day
Route	• Mode and route selection pattern between locations • Mode and route selection pattern between regions
Mode Choice	• Mode selection pattern between locations by mode • Mode selection pattern between regions by mode
Bus Operation and Management	• Number of boarding and disembarking, arrival and departure times by station, time period, route, and bus company

The first four types of traffic information outlined on the table above, excluding the contents on bus operation and management, are called the "Trip-Chain Data."

6.3 Characteristics of Data

The data collected through the T-Money system are the records of transit users with T-Money. The boarding and disembarking is recorded by card number. The data that is most useful for transportation planners is comprised of the trip-chain data, but only after further examination and research of the raw data that it provides. When the T-Money boarding/disembarking data is processed to the trip-chain data, the following issues were presented:

- Data of boarding records only: 90.3% of boarding/ disembarking records can be paired for bus records, and 100% for subway
- A record may be linked to a single transaction, but in reality it is composed of multiple departure and trips.
- A transfer transaction is only recognized within 30 minutes of modal transfer between transportation vehicles. Once exceeding the proposed time limit, the passenger transfer is recorded as two separate trips and the passenger will only receive a transaction discount.
- Records tagged before the disembarking station
- Cash user and fare-exempted user cannot be considered.
- Taking into consideration the trips of private automobile users and pedestrians
- Inter-regional long-distance bus users: Low percentage of correct records of corresponding boarding and disembarking

Alternative methods for the issues mentioned above are as follows, but remain under research.

- Develop an O-D matrix with only card users; excluding cash users, fare exempted users, private automobile users, and pedestrians.
- Derive the transaction sequence which is prepared by arranging the record of each card identification number by time sequence, ignoring the transaction units for the adjustment.
- Modify the records (MRecords) according to the four cases mentioned above. Then develop the necessary computer algorithm for the re-classification of the MRecords to be divided following the transaction unit (Modified Transaction Unit: MTU).
- Using the MRecord and re-filed MTU, derive the number of arriving passengers by origin, the number of arriving passengers by destination, the average travel time, distribution, and route, for each station and date.
- Develop the O-D matrix by every pair of boarding and disembarking at a station and time period of the day with the MRecord and the MTU.

7 Conclusion

The integrated transit-fare card system has been applied in Hong-Kong, Singapore, and Taiwan, but Seoul's integrated adjustment system, which includes the transaction discount, is the first trial in the world. With the reform of the bus-operating system, this integrated transit fare system increases the use of public transportation with an upgraded service quality of convenience at a discounted fare. It contributes to the decrease of traffic congestion by reducing the use of private automobiles, and by producing the raw data of travel demand and patterns within the metropolitan area, it provides a useful tool for the development of transportation planning, in addition to economic and environmental strategies.

However, it is necessary to develop a methodology to utilize the system not only for fare cards but also as a means for traffic data collection. The re-classification of modified records is a plausible suggestion, but the re-classification method for the record collected by T-Money should first be established logically and verified. The trips which do not use T-Money should also be a considered factor in the process of trip pattern analysis. The collection of modified data is better than the modification of collected data, as becomes more evident within a grouping of large amounts of data.

Currently, the traffic information generated from the T-Money system is applied to the bus route designation and distribution program. The T-Money system will be linked to the GIS databases of the city of Seoul, it's surrounding local governments, and the "National Geographic Information System." The data collected on the system will provide the basis for the trip demand and route demand analyses in the metropolitan area. The system will also be linked to the bus management system in the city of Seoul. The integrated transit-fare card system is not only an efficient arrangement with the easy-swipe method of a fare card, but it is also a new opportunity for development in the field of our urban transportation system.

References

1. Pucher, J., H. Park, M.H., Kim, J., Song, Public Transport in Seoul: Meeting the Burgeoning Travel Demands of a Megacity, Public Transport International, Vol. 54, No.3, 2005.
2. Ahn, K. and Y. Ohn, "Metropolitan Growth Management Policies in Seoul: A Critical Review" in Kwon, W. and Kim, K. editors, *Urban Management in Seoul: Policy Issues and Responses*. Seoul, South Korea: Seoul Development Institute, 2001, pp.49-72.
3. Hwang, K. "Transportation Policy in Seoul." in Kwon, W. and Kim, K. editors, *Urban Management in Seoul: Policy Issues and Responses*. Seoul, South Korea: Seoul Development Institute, 2001, pp.107-124.
4. Faghri, A. and Hua, J., Evaluation of Artificial Neural Network Applications in Transportation Engineering, Transportation Research Board 1358, 1991.
5. Haykin. S., Neural Networks-A Comprehensive Foundation. Prentice Hall, New Jersey, 1999.
6. Laurence C., and N. Tourigny, Road Safety Analysis: A Case-Based Reasoning Approach, Transportation Research Board, January 1998.
7. Introduction to e-Transportation Card System in Seoul Metropolitan City, Seoul Metropolitan Government, 2004.
8. Park, Y., The Data Processing System and the Derivative Traffic Information of T-Money, Anmin Forum, 2005.

Efficient Migration Scheme Using Backward Recovery Algorithm for Mobile Agents in WLAN

Dong Chun Lee

Dept. of Computer Science Howon Univ., Korea
ldch@sunny.howon.ac.kr

Abstract. We propose efficient migration algorithm with reordering and backward recovery of the paths to guarantee the migration of mobile agents in WLAN. The proposed method affords to avoid any faults of nodes or clients of mobile agents on WLAN.

1 Introduction

Mobile agent technology has been applied for various application areas such as wireless Internets management, and E-commerce [1, 3, 4, 7]. Mobile agents are autonomous objects that migrate from node to node of WLAN and provide to user which have executed themselves using database or computation resources of clients connected by the WLAN [3-8]. To migrate the mobile agent, it will be needed a virtual place so-called the mobile agent system to support mobility [8]. Several prototypes of mobile agent system have been proposed in several different agent systems such as Aglet [4], Agent TCL [9], Mole [2], and so forth.

Most systems are little ensured its migration for a fault of Internet nodes or a crash of clients (or hosts), which may happen during migrating after a mobile agent launch. On the faults such as a destruction of the nodes or the mobile agent systems, we may consider that mobile agents may be destroyed, or blocked against the seamless wired/wireless networks processing. It is no natural attribute to monitor the seamless progress of agent's execution, in a viewpoint to guarantee the autonomy of mobile agents. Therefore, we proposed a strategy with the simple techniques of path reordering and backward recovery to migrate mobile agents.

2 Related Work

While a mobile agent is launched to specific nodes/clients according to relevant routing schedules [4], it is possible to happen some problems about migration of mobile agent if the host happens an accident within where the agent visits and executes. Typically, ORB [6-7] implements distributed garbage collection in order to delete objects having no more references. Voyager [6] provides five policies for mobile agent's life cycle. Mole [2] supports the shadow protocol for orphan detection and successful termination for agents in mobile agent systems. The protocol is for detecting and processing what occur to any fault on migrating mobile agents. However, it does not provide to guarantee migration reliability of mobile agent. There is a simple protocol using transaction message queue, which is a procedure that the sender puts messages in the queue and receiver gets messages. There is also the same problem as

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 101–106, 2005.

the process of autonomous mobile agent in that it does not include facilities for monitoring the progress of an agent's execution. For example, assume that there is an agent in input queue of a host and the node's error occurs before the agent moves to queue of next node. Then the agent is blocked until that the node is recovered. This situation differs from problem in client/server. Mole [2] provides a fault tolerant protocol to support effective way for 'exactly once' migration using voting and selecting protocol as copying mobile agent to all nodes.

3 Proposed Migration Model

Figure 1 show how the node repository can use in implementation instead of transaction message queue for agents. Assume that an agent moves from a node to the consecutive node along the path N1→N2→ … →N (k-1) →Nk (where Ni is a Internet node, Hi is a client (or a mobile host), and Ri is an agent repository). As an agent may visit the same node several times Ni and Nj (1<=i, j<=k) may denote the same or different nodes. Assume further that an agent is stored in a repository when it is accepted by the agent system for execution. Except Nk, each other node performs the following sequence of operations on Transaction Ti such as Get (agent); Execute (agent); Put (agent); Commit.

Get removes an agent from the node's repository. *Execute* performs the received agent locally. *Put* places it on the repository of the host that will be visited the right next time. Three operations are performed within a transaction and hence consisted of the atomic unit of work.

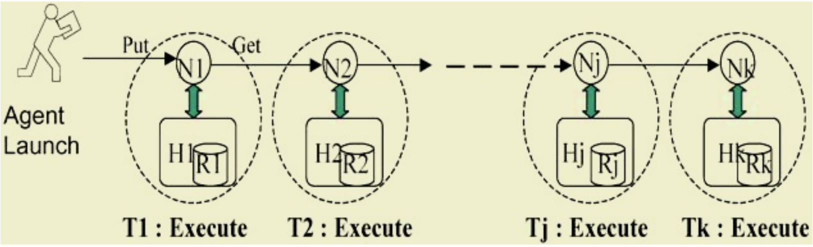


Fig. 1. Migration path of a mobile agent

In Figure 1, we assume to happen to a failure in a particular node Ni within the migration path of the mobile agent. For example, although the node Ni of the client H_i lives, the agent can't be migrated. Inversely, although the node Ni can be communicated with the previous node N (i-1), the agent can't occasionally migrate if the client Hi does not operate the agent system. In the above cases, the agent is never arrived by the last node Nk. It may be very weak point that the agent at previous client Hi-1 needs to receive user's assertion. In the worse case, if a shared client on the multiple agents launched occurs to crash on launching, the agents will block or destroy even if other nodes are available to process continuously. Therefore, it is ineffective to the mobility of agents. To solve these situations, we propose backward recovery algorithm of reliable migration for mobile agent in WLAN.

In Figure 2, we suppose that the migrated agents execute autonomously at the client H5. If the client H5 of node N5 crashes, all agents at that host are blocked or destroyed. To prevent it, the agent copies itself (that is, the clone) in the current client when an agent migrates the next node after it ends its job at the current client. The clone is unconditionally waiting until receiving an acknowledge signal 'ACK' from the next client. If the signal 'ACK' doesn't reach to the current client H4 within the timestamp from the next client H5, the cloned agent at the client H4 has automatically activated to resolve this hindrance.

Consequently, it hops to the next node N6. If the agent faults at the client H6 on execution, it will work repeatedly the same method to connect the other next node.

In Figure 3, it happens that the current running agent at a client H5 is destroyed by a particular clash. At the same time, if the current client H4 also happens to the succeeding fault, a cloned agent (which has already copied the previous client H3 of the client H4) wakes up and re-runs. This is so-called Backward Recovery.

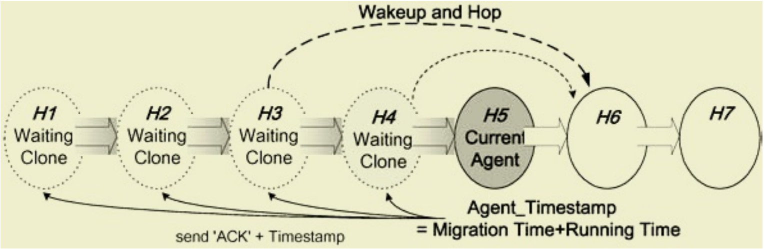


Fig. 2. Backward Recovery

The backward recovery method is as follows: the agent system copies the clones of its agent on the client before migrating to the right next client. Each clone is waiting by it's own timestamp. Its timestamp of the original clone is maximum at the original host of the migration path, the next clone will be less than the migration and execution time of the previous one, and so forth. From launching an agent, the timestamp accumulates every clone of the previous clients with it's own moving and running time before it depart for the current client. Therefore, clones are waiting by the timestamp. Each clone spontaneously revives and attempts to work the path reordering as soon as regarding as a clashed host when none received any signal from the next host. At the final destination's node of the host, the agent system should broadcast a signal 'Agent_Fire' to all copied clones of the agent excepting the faulty nodes and failed hosts until reaching the destination.

In such method, the algorithm provides a backward recovery method for which mobile agents support efficient migration from a host to the next one.

Algorithm Backward Recovery

```
Waiting Clones Check
// Periodically checking the timestamps of the clones.
for each sleeped_clone
   if (empty a clone_timestamp) {
     Notify to user;
     Call wakeup Clone;
    }
```

```
Go Agent  // migrating the agent or cloning it.
  Send the agent;
  Wait the agent's 'ACK' signal during send_timestamp;
  if ('ACK') {
       Clone the agent;
       Call sleepAgent;
  }
  else call wakeupClone;

Arrive Agent  // if the agent arrives in the JAMAS, noticing the previous node with the signal and
          //and executing.
  Send 'ACK' to the previous_node;
  Execute the agent;

Sleep Agent  // Each cloned agent is waiting for the assigned timestamps
 for each cloned_agent {
   Add agent_timstamp to system_timestamp;
   Add the agent to the sleeped_list;
   Sleep the agent;
 }

Wakeup Clone  // re-activating the clone of the agent.
 for the sleeped_list
      Find a cloned_agent;
 Remove it from the sleep_list;
 if ('Agent_Fire') remove the cloned_agent;
 else  {
       Move the current failed_address to last in the routing- table;
       Set the fail_checked information;
       Call the reordering algorithm; // arranging the path.
      }
```

4 Implementation

The proposed system consists of Graphic User Interface, Agents Mobile Service Component, Agents Execution Environment Component, and Agents Repository to provide the naming transparency of agents. In addition, it may be executing one more systems within a host. We show to launch through the process of an agent which manages some NEs (Internet elements). The following figures show that the simple agent as a role of Management Information Base (MIB) browser should be migrated and executed according to the routing schedule.

Figure 3 depicts the routing path of the agent such as $NE_h \rightarrow NE_b \rightarrow NE_a \rightarrow NE_c$, and we assume to be a fault at the client NE_b.

The Internet manager fetches the prepared agent and specifies routing addresses of it to migrate. So, clicking the ‘Go’ button on the manager’s window to launch it, the agent starts on a tour to get the MIB information of each NE on behalf of the WLAN manager.

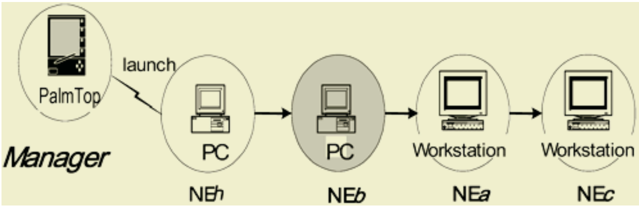
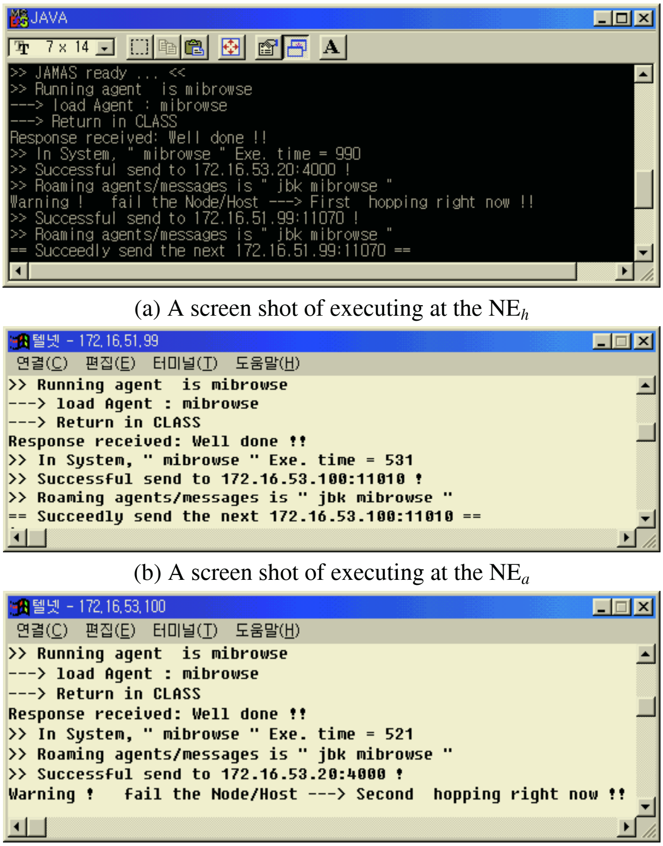


Fig. 3. A routing path with a fault of clients NE_b

Figure 4 shows executions of the agent at each NE as follows: Figure 4 (a), as a screen capture of the host NE_h, shows hopping by a failure of connection at the next NE_b after the launched agent normally progresses. Due to a failure of the host, the agent passes to next one. Thereafter, Figure 4(b), (c) capture executing of the agent at the hosts NE*a* and NE*c*. Then it is adapted to the proposed method. Therefore, the agent has toured for all nodes having no faults before that it does re-connect with the faulty nodes.



```
>> JAMAS ready ... <<
>> Running agent  is mibrowse
---> load Agent : mibrowse
---> Return in CLASS
Response received: Well done !!
>> In System, " mibrowse " Exe. time = 990
>> Successful send to 172.16.53.20:4000 !
>> Roaming agents/messages is " jbk mibrowse "
Warning !   fail the Node/Host ---> First  hopping right now !!
>> Successful send to 172.16.51.99:11070 !
>> Roaming agents/messages is " jbk mibrowse "
== Succeedly send the next 172.16.51.99:11070 ==
```

(a) A screen shot of executing at the NE_h

```
>> Running agent  is mibrowse
---> load Agent : mibrowse
---> Return in CLASS
Response received: Well done !!
>> In System, " mibrowse " Exe. time = 531
>> Successful send to 172.16.53.100:11010 !
>> Roaming agents/messages is " jbk mibrowse "
== Succeedly send the next 172.16.53.100:11010 ==
```

(b) A screen shot of executing at the NE_a

```
>> Running agent  is mibrowse
---> load Agent : mibrowse
---> Return in CLASS
Response received: Well done !!
>> In System, " mibrowse " Exe. time = 521
>> Roaming agents/messages is " jbk mibrowse "
>> Successful send to 172.16.53.20:4000 !
Warning !   fail the Node/Host ---> Second  hopping right now !!
```

(c) A screen shot of executing at the NE*c* and attempting migration of the second at the NE_b

Fig. 4. Fault-tolerable executions of a mobile agent at each NE

Therefore, the efficient migration scheme for mobile agents ensures the persistency of computation to preserve autonomous mobility and information of state for agents though there are some faults of nodes or clients on the routing schedules of the WLAN.

5 Conclusion

In this paper we introduce the backward recovery algorithm to ensure the migration of mobile agents in WLAN. The proposed algorithm not only affords to avoid any faults of nodes or clients of mobile agents on WLAN but also affects to agents' life span. All presented techniques have been implemented in our system. Therefore, our system can improve effectively the problem of performance and WLAN overhead due to the imposed characteristics of distributed architecture since a mobile agent offers not only the migration reliability and transparency for mobile agent as autonomously as possible but also computing environment which is capable of distributed processing with mobile objects. Future work will investigate for the agent groups with distributed event services on WLAN.

Acknowledgements

This work was supported by the Howon University Fund, 2005

References

1. K.A. Baharat, L. Cardelli, "Migratory Applications", Proc. of the 8th Annual ACM Symp. on UISTech., November 1998.
2. J.Baumann, " A Protocol for Orphan Detection and Termination in Mobile Agent Systems", TR-1997-09, Stuttgart Univ. Jul., 1999.
3. General Magic, "Odyssey", URL: http://www.genmagic.com/agents/
4. IBM, "The Aglets Workbench", URL: http: //www.trl.ibm.co.jp/aglets
5. D. B. Lange, M. Oshima, "Seven good reasons for mobile agents", Proc. of CACM, Vol. 42(3), Mar. 2002, PP 88-89.
6. Objectspace Voyager, GeneralMagic Odyssey, IBM Aglets: A Comparison, June, 2002.
7. OMG, "Mobile Agent Facility Interoperability Facilities Specification (MAF)", OMG.
8. A. Puliafito et al., "A Java-based Distributed Network Management Architecture", 3rd Int'l Conf. on Computer Science and Informatics (CS&I'99), Mar. 1999.
9. Robert S.G., "AgentTCL: A Flexible and Secure Mobile-agent System", TR98-327, Dartmouth Col. June 1999.

Using Similarity Measure to Enhance the Robustness of Web Access Prediction Model

Ben Niu and Simon C.K. Shiu

Department of Computing, Hong Kong Polytechnic University, Hong Kong, China
{csniuben,csckshiu}@comp.polyu.edu.hk

Abstract. Prefetching web content by predicting users' web requests can reduce the response time of the web server and optimize the network traffic. The Markov model that is based on the conditional probability has been studied by many researchers for web access path prediction. The prediction accuracy rate can reach up to 60 to 70 percent high. However a drawback of this type of model is that as the length of the access path grows the chance of successful path matching will decrease and the model will become inapplicable. In order to preserving the applicability as well as improving the accuracy rate, we extend the model by introducing a similarity measure among access paths. Therefore, the matching process becomes less rigid and the model will be more applicable and robust to the change of the path length.

1 Introduction

Web users nowadays are becoming more impatient and may not wait at all if the time for downloading a web page takes more than a few seconds. Furthermore, many web pages contain a substantial amount of huge size multimedia data, which takes a great deal of time in disk seeking, I/O transmission and Internet transport. In order to reduce the web access latency (i.e., response time), the demand of effectively caching the web content is increasing.

For prefetching task the Markov model which is based on conditional probability has been studied by many researchers [1]-[7] before. The prediction accuracy rate of this type of model can reach up to 60 to 70 percent high. It was also found that as the order of the Markov model increases the accuracy rate can also be increased linearly. But this increase in accuracy rate of prediction is at the cost of degrading the model's applicability because long web access paths are rare and it will be more difficult to get successful pattern matching as the length of the paths increases. We proposed a new approach that is based on the similarity measure among access paths to model the user access behavior so that the effect of the tradeoff between accuracy and applicability can be minimized. Thus, the model will be more robust to the change of the path length.

2 Related Work

The previous solutions for web access path prediction are mostly based on the kth-order Markov model. In the discrete-time case, the model assumes that the conditional probability of a system's next state depends on a finite history of its previous states, as formalized in formula (1),

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 107–111, 2005.

$$P(S_{i+k} \mid S_i, S_{i+1}, \ldots, S_{i+k-1}) = \\ P(S_{i+k}, S_i, S_{i+1}, \ldots, S_{i+k-1}) / P(S_i, S_{i+1}, \ldots, S_{i+k-1}) \quad (1)$$

where S_i denotes the state of a system at time i. The model is called the first-order Markov model when k equals to one.

3 Using Similarity to Improve the Robustness of Web Access Prediction Model

The similarity measure is used for user group clustering and prediction support value calculation.

3.1 Measuring the Similarity of Web Access Paths

Each access path in the access log file can be represented as a string with each accessed web element such as a page, an image or a piece of CGI code denoted as a character. The similarity between any two web access paths is measured by checking the Levenshtein distance, a basic form of the edit distance, between the two corresponding strings. The distance value is computed by counting the operation costs of transforming one string into the other. They have been successfully used in spell checking, speech recognition, DNA analysis, and information retrieval where approximate matching of string pattern is required.

3.2 Clustering User Groups Using the Access Path Similarity

The users are grouped by clustering the web access paths since their access history reflects their interests while browsing the web sites. We choose the Group Average Agglomerative Clustering method for this task. Let $L = \{S_1, S_2, \ldots, S_r\}$ be the set of the strings representing the access paths. The clustering algorithm starts by checking the similarity between each pair of points in L. The most similar pair(s) will be put into the same cluster(s). Recursively the most similar clusters will be further merged to generate new clusters. For two clusters $C_i, C_j \subseteq L$ the inter-cluster similarity CSim is defined as,

$$\mathrm{CSim}(C_i, C_j) = \frac{\sum_{Sl \in Ci} \sum_{Sm \in Cj} \mathrm{sim}(S_l, S_m)}{|C_i| \cdot |C_j|} \quad (2)$$

After merging for several times there will be only one remaining cluster and a hierarchical tree is generated. Each branching point in the tree has a similarity score showing how similar the clusters under that branch are to each other. By clustering the web users are put into different groups. The prediction is performed using the path information from a single user group rather than the whole user profile. This reduces the computation cost and keeps the prediction process from being interfered by the noise from other user groups. The prediction is thus made more robust and efficient.

3.3 Calculating the Support Value of Prediction Using Similarity and Conditional Probability

The model uses the similarity measure and probability to make predictions. Given a pattern path the model first checks out which user group it belongs to by measuring the distance between the path and the clusters. In the identified group, it calculates its similarity to the existing paths that contains the last accessing point in the pattern path. The support value to jump from the last accessing point to the next point is computed with the following formula,

$$\begin{gathered} sp(D_0, h_l) = \sum_{i=1}^{n} sim(D_0, D_i) \cdot P(h_l \mid D_i) \\ P(h_l \mid D_i) = P(D_i, h_l) / P(D_i) \end{gathered} \tag{3}$$

where D_0 is the given pattern path, h_l is one of the possible destinations in the next hop from the last accessing point in D_0. $sp(D_0, h_l)$ is the support value of going from D_0 to h_l. D_i is the ith access path in the user group identified and it contain the last accessing point of D_0. The conditional probability $P(h_l \mid D_i)$ describes the likelihood of going to h_l from D_i. It is clear that in the above equation each qualified path in the user group contributes to the support value of the predication. The contribution of a path is in proportional to its similarity to the pattern path.

4 Example

In this section, an illustrative example is given. Suppose after performing clustering we have three user groups.

Group 1: A B C E F
A B C E F
A B C E D
A B C D E F
Group 2: B C D K A
B C D K B
B C E K B
Group 3: B C E F K
B F C E D

Each line of string represents an access path and each character represents a web element accessed by users. Given a pattern path A C E with the last accessing point being E the problems is to predict the next web element that user will most likely request. However, we find no complete matching in the above three groups. The previous method handles this by just simply decreasing the order of gram. Therefore A C E will be shortened to C E for further matching. The drawback of doing so is that noise will be introduced from other user groups into the prediction process in the current group. As we can see that path A C E belongs to group 1 as its similarity to group 1 is the largest, while C E occurs in not only group 1 but also in group 2 and group 3 in which users have different motivations in navigation. If we use the path information in group 2 and group 3 to make predictions for the users in group 1 the accuracy rate of prediction will be affected. In the example if we use the original method to make prediction, the result will be as follows,

Table 1. Prediction result using Markov model

3-Gram pattern path	Prediction with support value
A C E	F (3/6 = 50%)
A C E	D (2/6 ≈ 33%)
A C E	K (1/6 ≈ 17%)

Using the similarity based method we have different result. First, calculate the similarity between the pattern path and all the other paths in each group. It is found that ACE should belong to group 1. The prediction for this pattern path will thus be based on the information from this group. Therefore by calculating the similarities of the strings and the probabilities,

$$\mathrm{Sim(ACE,ABCE)} = 1/(1+1) = 0.5$$

$$\mathrm{Sim(ACE,ABCDE)} = 1/(1+2) \approx 0.33$$

$$\mathrm{P}(\{\mathrm{F}\} \mid \{\mathrm{ABCE}\}) = 2/3 \approx 0.67$$

$$\mathrm{P}(\{\mathrm{F}\} \mid \{\mathrm{ABCDE}\}) = 1/1 = 1$$

$$\mathrm{P}(\{\mathrm{D}\} \mid \{\mathrm{ABCE}\}) = 1/3 \approx 0.33$$

Using formula (3) we obtain the probabilities of arriving at F and D in the next hop,

$$\begin{aligned}&\mathrm{sp(ACE,F)}\\ &\quad = \mathrm{Sim(ACE,ABCE)} \cdot \mathrm{P}(\{\mathrm{F}\} \mid \{\mathrm{ABCE}\}) +\\ &\qquad \mathrm{Sim(ACE,ABCDE)} \cdot \mathrm{P}(\{\mathrm{F}\} \mid \{\mathrm{ABCDE}\})\\ &\quad = 0.5 \cdot 0.67 + 0.33 \cdot 1\\ &\quad = 0.665\end{aligned}$$

$$\begin{aligned}&\mathrm{sp(ACE,D)}\\ &\quad = \mathrm{Sim(ACE,ABCE)} \cdot \mathrm{P}(\{\mathrm{D}\} \mid \{\mathrm{ABCE}\})\\ &\quad = 0.5 \cdot 0.33\\ &\quad = 0.165\end{aligned}$$

It should be noted that a performance tradeoff between the accuracy and the efficiency exists because the given pattern paths may be clustered into wrong user groups, which will decrease the prediction accuracy. In this case controlling parameters such as the level of clusters should be adjusted according to the application demands.

5 Conclusions

In this paper we introduce the similarity measure into the web access prediction model to enhance its robustness for making predictions. The similarity measure is first used to group the web users by referring to their past accessing logs. The prediction is made based on the historical records in a single user group. This reduces the possibility for the prediction being interfered by the records in other user groups. Also, by using the similarity based approximate pattern matching the chance of mismatching between the pattern path and the historical path can be reduced which further improve the robustness of the model.

References

1. Griffioen, J., Appleton, R.: Reducing file system latency using a predictive approach, Proc. USENIX Conference, Boston, Massachusetts, USA (1994) 8-12
2. Padmanabhan, V.N., Mogul, J.C.: Using predictive prefetching to improve World Wide Web latency, ACM Computer Communication Review, Vol. 27, no.3, (1996) 22-36
3. Pitkow, J., Pirolli, P.: Mining longest repeated subsequences to predict World Wide Web surfing, Second USENIX Symposium on Internet Technologies and Systems, Boulder, Colorado, USA, (1999) 11-14
4. Su, Z., Yang, Q., Lu, Y., and Zhang, H.: WhatNext: A prediction system for web request using N-gram sequence models, First International Conference on Web Information Systems and Engineering Conference, Hong Kong, China (2000) 214-221
5. Yang, Q., Zhang, H., and Li, I.: Mining web logs for prediction models in WWW caching and prefetching, The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD'01, San Francisco, California, USA (2001) 473-478
6. Yang, Q., Zhang, H.: Web log mining for predictive web caching, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, no. 4, (2003) 1050-1053
7. Pal, S. K., Shiu S. C. K., Foundations of Soft Case-Based Reasoning, John Wiley, Hoboken, New Jersey, 2004

Intelligent Financial News Digest System

James N.K. Liu[1], Honghua Dai[2], and Lina Zhou[3]

[1] Department of Computing, Hong Kong Polytechnic University
[2] School of Information Technology, Deakin University, Australia
[3] Department of Information Systems, University of Maryland, Baltimore County, USA

Abstract. We present an agent-based system Intelligent Financial News Digest System (IFNDS) for analyzing online financial news articles and associated material. The system can abstract, synthesize, digest, and classify the contents, and assesses whether the report is favorable to any company discussed in the reports. It integrates artificial intelligence technologies including traditional information retrieval and extraction techniques for the news analysis. It makes use of keyword statistics and backpropagation training data to identify companies named in reportage whether it is, evaluatively speaking, positive, negative or neutral. The system would be of use to media such as clipping services, media management, advertising, public relations, public interest, and e-commerce professionals and government non-governmental bodies interested in monitoring the media profiles of corporations, products, and issues.

1 Introduction

The Web offers access to vast amounts of information, but the usefulness of this access is limited by our ability to make sense of it in a timely way. Our ability to locate, filter, and analyse this information can be much improved by making use of intelligent techniques and agent technologies [13]. One example of this might be the processing of quantitative data by automatic computer programs using artificial intelligence techniques such as fuzzy logic for knowledge representation and for making useful inferences or actions; expert systems for evidential and heuristic reasoning [5]; neural networks for classification and adaptive learning [6]; genetic algorithms for solution optimisation [13]; and data mining techniques for knowledge discovery [7]. Similarly, agent technologies now allow us to build software modules that can monitor, assist, and act on behalf of a user and to inter-operate with other agents. Such technologies have also considerable potential for information retrieval.

Recent research in information retrieval, information extraction, and text categorization has produced data classifications that are for many purposes too broad [8, 9, 12, 14, 15, 16, 17, 18, 20] and that as a result have limited explanatory or analytical power. For example, a simple statistic such as the frequency of occurrence of a company name by itself tells us nothing of the tenor of the report, whether the reportage reflects positively or negatively on the company. We call this the "insight" of the news and touch upon natural language processing areas. Some quantitative measure of it will be discussed in the paper.

FIDS [11] classified company news into thematic categories such as company performance, economy, mergers, products, and securities. SCISOR [9] was much more

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 112–120, 2005.

narrowly focused, having just one theme, corporate takeovers. In both cases, a reader would require substantial domain knowledge and time in order to arrive at a decision as to the connotation of a report. In this paper, we propose an Intelligent Financial News Digest System (IFNDS), that uses a simple network design and keyword searching mechanism not just to efficiently extract and classify news articles but which will also quickly classify the connotation of a report along a scale of positive to negative view independent of the theme, context, or domain of the report.

The primary goal of this research is to improve the speed and accuracy and usefulness to the end user of information retrieval in the context of daily financial news. The vast volume of news makes categorization the first task. In this research, we limit ourselves to the financial news pages of WordNet [1, 2]. We approach the primary issue of automatic text abstraction or summarization [14, 17] and by first selecting a document based on its salient features, such as its theme, location, title and a number of assigned cue features that are associated with the evaluative stance of the reportage. We then analyze the document according to its hierarchical structure: chapters, sections, subsections, paragraphs, sentences, terms, words and characters. Sentences are then selected from the source document based on an assigned weighting for their significance in the document which is derived by computing the weighted sum of the weights of all the features. Sentences with sentence weights higher than a threshold value are selected to form part of the summary.

2 System Overview

Figure 1 provides an overview of IFINDS which is composed of three agents - NewsAglet; CompanyIdentifer; and InsightClassificationAglet – and an enquiry interface, the WebController servlet. It operates as follows: First, financial news articles are, according to a schedule, automatically downloaded from the Internet and stored in local host or remote site. Keywords are then loaded from a database to NewsAglet which uses keyword matching to extract sentences that contain or refer to a nominated term (e.g. company name) and perform statistical analyses relevant to the keyword. Extracted sentences and keyword analyses are sent to the WebController which passes the keyword statistics to CompanyIdentifier and the statistics on positive and negative keywords to the InsightClassificationAglet. The CompanyIdentifier queries a database to convert the keywords into a stock code for a particular company and the InsightClassificationAglet classifies the news using knowledge from a trained back propagation neural network. The InsightClassificationAglet sends the results of insight classification to WebController, which gathers and displays the results from the three agents.

3 Data Collection

This section describes the IFNDS data collection methodology. It involves the collection of documents via shallow text analysis, the three-step process in which NewsAglet, the keyword matching and extraction agent, scans at the sentence and sub-sentence level in order to identify the company under discussion in a news report, and the associated keyword loading construct.

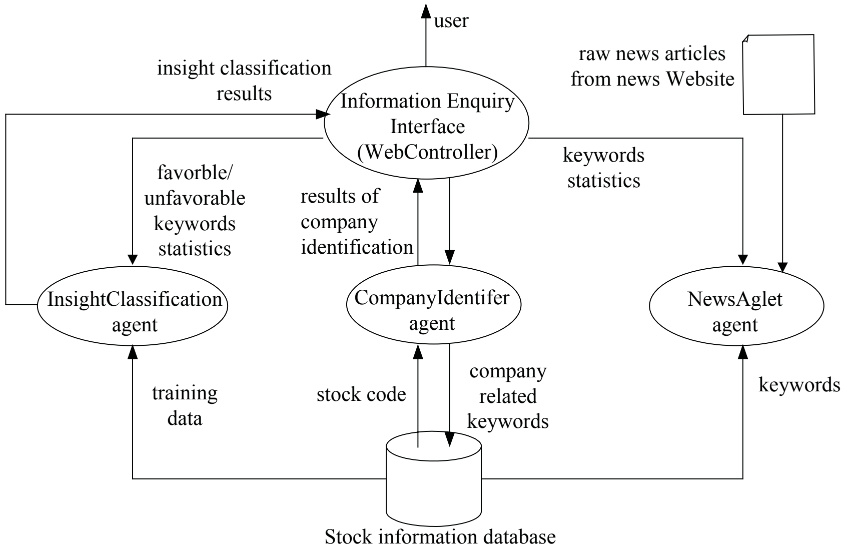

user
insight classification results
Information Enquiry Interface (WebController)
raw news articles from news Website
favorble/ unfavorable keywords statistics
results of company identification
keywords statistics
InsightClassification agent
CompanyIdentifer agent
NewsAglet agent
training data
stock code
company related keywords
keywords
Stock information database

Fig. 1. An overview of the IFNDS system

3.1 Shallow Text Analysis

Shallow text analysis can reduce resource-consuming tasks by performing relatively shallow text analysis and to find related documents to pinpoint the exact result, but all these take time. IFNDS makes no attempt to do approximate matching and return the sentences with the highest semantic content overlaps. If the process fails, the system attempts keyword matching [e.g. 14], in which it abandons syntactic criteria and users only collect information about word classes. A traditional passage retrieval that is enhanced with parts-of-speech tags will then be resembled.

3.2 Scanning of the News Articles by NewsAglet

NewsAglet identifies the company under discussion in a news report by first processing sentences, then individual words, and finally storing these as word tokens and sentence indexes. We create a small thesaurus for WordNet and a number of other news sites, and also define synonym identifiers for all of the thesaurus's synonym sets. When CompanyIdentifier receives the results of keyword statistics, it will convert each keyword to the related stock code of the listed company by querying the local database. It then will accumulate the stock code frequency and find out which stock code occurrence frequency is the highest. The company, which has the highest occurrence frequency, will be identified as the target company reported in the news. The details are as follows.

Processing Sentences in News

In order to increase the efficiency of encoding the news articles, when the news article is firstly read by NewsAglet, punctuation marks like (!, ?) will be replaced by a full stop with 2 white spaces (.) in the news contents. Then all occurrences of full stop will be tokenized into sentences. Each sentence ***S*** will be stored in NewsAglet by its sequence order in the news article.

sentences = (S_1, S_2, S_3, S_4, S_5..... S_i) where i = sentence sequence index

Processing Words in News
After processing sentences, all punctuation marks except full stop will be replaced with white space in each sentence. This is to avoid mistake occurred when these punctuation marks concatenate the keyword when keyword matching is being processed.

Storing Word Tokens of News in NewsAglet
We store both the sentence index *i* and word token to provide the word location for NewsAglet to perform sentence extraction after the matching of those keywords.

3.3 Loading Keywords for Matching in NewsAglet

All keywords or keyword phrases are loaded into some object class. The keywords class is a hash table, which has 26 records. Each record has a unique alphabet (a – z) as the key and since all keywords are divided into 26 entries in the hash table, the number of comparisons between keywords and the words in the news articles can then be reduced accordingly.

4 Insight Classification Scheme

In order to set up the standard for classifying whether the news is favorable or unfavorable to a listed company, we have set up an insight classification scheme that can be used to classify news articles (see Table 1) as a first step in assembling training data. This information is then used in the training of a back propagation network. Each class of the news corresponds to a range of rating scores as shown in Table 2. For the back propagation network to learn how to classify whether the news is favorable to the company mainly reported, we tried to use the number of favorable keywords and the non-favorable keywords in / not in the abstract as the feature input:

1. No. of words belonging to favor_word table in the abstract
2. No. of words belonging to favor_word table not in the abstract
3. No. of words belonging to non_favor_word table in the abstract
4. No. of words belonging to non_favor_word table not in the abstract

The abstract will contain keywords of the listed company in the report. Therefore, the favorable and non-favorable keywords should have more effect than those keywords not in the abstract.

After users have finished classifying the news articles, the results of class rating will be converted to some rating score which is the median of the range of each rating class respectively. The inputs and news rating of the news will be input into database as the training data, which will be used to train the back propagation network.

Table 1. 5-class-rating of insight classification

Rating	Descriptions
very good	• the news mainly contains lots of comment / analysis / information which are favorable to the company
good	• the news contains more favorable than non-favorable comment / analysis / information about the listed company
neutral	• the news does not contain significant comment / analysis / information about the company
bad	• the news contains comment / analysis / information which is unfavorable to the company
very bad	• the news contains more non-favorable then favorable comment / analysis / information about the company

Table 2. Rating scores used in back propagation network

Class field	Rating score range in back propagation network
very good	0.8000 - 1.0000
good	0.6000 - 0.7999
neutral	0.4000 - 0.5999
bad	0.2000 - 0.3999
very bad	0.0000 - 0.1999

5 Experimental Results

To evaluate the performance of the system, a total of 378 news articles were collected from the web site of the Hong Kong Standard (www.hkstandard.com.hk), an English-language newspaper published in Hong Kong. These articles were used to test the performance of IFNDS in identifying news that substantially comments on a particular company and in classifying the reportage in terms of the positive-negative classification scheme. The system test results are compared with results obtained by a human rater operating upon the same news material. Typical results include:

Test 1 – Identifying the Relevant Listed Companies in News Articles (Extraction of Sentences Related to Listed Companies)

As the extraction of sentences in news articles is dependent on the identification of company related keywords (company alias, company person, major shareholder), in order to measure the performance of extraction effectiveness, we have to evaluate the performance of company keyword identification. Since the identification of relevant listed companies does not require any training, we used all 378 news articles gathered to test the performance. These are defined as follows:

$$\text{Recall} = \frac{\text{no. of news articles have all the relevant listed companies correctly identified}}{\text{Total no. of news articles}}$$

$$\text{Precision} = \frac{\text{no. of news articles have all the relevant listed companies correctly identified}}{\text{Total no. of news articles extracted}}$$

The system has an overall precision of 94.38% and recall of 64.28% in the test of identifying relevant listed companies in news articles (Table 3). The reason of having a low recall rate is that when gathering our news articles, we downloaded many financial news articles which did not refer to listed companies. There were quite a number of financial news concerning the local economy and financial policy of government. Moreover, the data collection period coincided with the Hong Kong government announcement of its budget, many news reported the responses from the community such as "Lawmakers united in opposition to tax increases"; "Budget lesser of two evils for middle class".

Table 3. Test results of identifying relevant listed companies in the news articles

Items	Extracted	Correct	Incorrect	Precision	Recall
378	243	229	14	0.9438	0.6428

Test 2 – Identifying the Listed Company Mainly Reported in News Articles

We used the same 378 news articles to test the performance. The measures for the effectiveness of the identification of the listed company mainly reported in news article are:

$$\text{Recall} = \frac{\text{no. of news articles correctly identified for the listed company mainly reported in the news}}{\text{Total no. of news articles}}$$

$$\text{Precision} = \frac{\text{no. of news articles correctly identified for the listed company mainly reported in the news}}{\text{Total no. of news articles extracted}}$$

The test results are shown in Table 4. The system has an overall precision of 81.46% and recall of 64.28%. Compared with the precision rate in the previous test, it dropped to 81.46%, which is somehow unexpected. We discovered that a lot of news quoted the comment of some economists from a few famous banks like "HSBC", "Hang Seng Bank" and these alias names occurred even more frequently in many news reports. Although the economists are employed by these banks, the news articles certainly did not report much about these banks.

Table 4. Test results of identifying the listed company mainly reported in the news

Items	Extracted	Correct	Incorrect	Precision	Recall
378	243	197	46	0.8146	0.6428

Test 3 – Classifying News Insight

Preparation

Users have classified whether the news is very bad, bad, neutral, good or very good to the listed company mainly reported according to the classification scheme mentioned earlier. From the total of 378 news articles collected, only 197 news articles were successfully identified as ones with news insight reported. We then split these 197 news articles into two sets, a training set and a testing set. The training set contains 127 articles and the testing set contains 70 articles. Table 5 shows the distribution of news articles in each rating class.

Table 5. Number of news articles in each rating classes in the training data set

Rating class	No. of news articles
very good to the company mainly reported in the news	17
good to the company mainly reported in the news	35
neutral to the company mainly reported in the news	41
bad to the company mainly reported in the news	26
very bad to the company mainly reported in the news	8

Measure

The measure for the effectiveness of the news insight classification is defined as follows:

$$\text{Precision} = \frac{\text{no. of news articles correctly classified per news rating class}}{\text{Total no. of news articles classified corresponding to the news rating class}}$$

In the test of classifying news insight, the system has attained an overall precision of 68.57%. The overall results remain to be improved, but we are glad that the performance for classifying news of class "very bad" and "very good" have achieved much better precision rate of 90% and 89.47% respectively.

In evaluating the error made in the classes "bad", and "neutral", we found that most of the classifying mistakes were in classifying news as "very good". This trend appeared to be abnormal because even in the distribution of news articles in the training set, the number of news articles classified as "neutral" was the highest. Also, it is easily observed that the precision rate of the classes "bad", "neutral", "good" are significantly lower than the classes "very bad" and "very good". A summary of the test results is shown in Table 6.

Table 6. Test result summary of classifying news insight

	Actual Rating (no. of files in rating)				
Predicted Rating (no. of files in rating)	**very bad (10)**	**bad (15)**	**neutral (12)**	**good (14)**	**very good (19)**
very bad (11)	9	0	0	2	0
bad (12)	1	8	0	2	1
neutral (5)	0	0	5	0	0
good (10)	0	0	0	9	1
very good (32)	0	7	7	1	17
0Accuracy	90%	53.33%	41.67%	64.29%	89.47%

6 Conclusion and Future Work

We have presented an intelligent system for analyzing financial news on Web. It incorporates with neural network and agent technologies to help extract and classify news articles. With the reference of WordNet information and keyword statistics, the proposed system is able to identify news reporting favorably or non-favorably to a given company. For future work, we plan to optimize the system performance by improving the extraction scheme. The current system can only process listed company related news articles, therefore we would like to further develop the system to handle news in different domains and of different types, and to improve the recall rate of the news extraction. Moreover, as the insight classification is really a new and challenging task which can be difficult to assess due to subjective and imprecise human judgment, we need to explore the investigation by enriching the knowledge representation and make the system adaptive so that it can learn the news insight better.. We are also in the process of extending the system to handle multi-lingual news from different websites. Techniques over natural language processing to extend the shallow text analysis to semantic understanding of the news content will be explored later on.

Acknowledgement

The authors would like to acknowledge partial support through the ICRG grant A-PF83 of The Hong Kong Polytechnic University.

References

1. http://wordnet.com.au
2. http://learn.tsinghua.edu.cn/homepage/2001315450/wordnet.html
3. Java Servlet 2.1, http://java.sun.com/products/servlet/2.1/
4. Costantino, M. (1999), "IE-Expert: Integrating natural language processing and expert system techniques for real-time equity derivatives trading", Journal of Computational Intelligence in Finance, 7(2), pp. 34-52.
5. Dai, H. and Wang, J. (2000) WWW mining and intelligent Internet decision aid, *International Conference on Modelling and Simulation*, IASTED, Canada
6. Finkeistein-Landau, M. and Morin, E. (1999), "Extracting semantic relationships between terms: Supervised vs. unsupervised methods", in Proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure, Dagstuhi Castle, Germany, May, 71-80.
7. Hang Li, Yunbo Cao, and Cong Li (2003) "Using bilingual Web data to mine and rank translations", IEEE Intelligence Systems, 18(4), pp. 54-59
8. Hans van Halteren (2003) "New feature sets for summarization by sentence extraction", IEEE Intelligence Systems, 18(4), pp. 34-42.
9. Pauls, Jacobs, Lisa F. Rau, SCISOR: Extracting Information from online News, CACM Vol. 33, No. 11, pp 88-97, 1990
10. Jun-Tae Kim and Dan I. Moldovan, "Acquisition of Semantic Patterns for Information Extraction from Corpora", IEEE, 1993
11. Wai Lam and Kei Shiu Ho (2001) "FIDS: An Intelligent Financial Web News Articles Digest System", IEEE Trans. On SMC, 31(6).
12. Li, Wenjie, Wong, KamFai and Yuan, ChunFa (2003), "A design of temporal event extraction from Chinese financial news", International Journal of Computer and Chinese Language Computer Society, 16(1), pp. 21-39.
13. Liu, James N.K., Kwong, Raymond W.M. and You, J (2002), "Towards an intelligent Web-based agent system (iWAF) for e-finance application", in *Proceedings of the IASTED International Conference on Artificial and Computational Intelligence, September 25-27, 2002, Tokyo, Japan, pp. 18-23.*
14. Liu, J., Zhou, L. and Wu, Y. (1999), "A hybrid method for abstracting newspaper articles", Journal of the American Society for Information Science, 50(13), pp. 1234-1245.

15. Diego Molla, Rolf Schwitter, Fabio Rinaldi, James Dowdall and Michael Hess (2003) "ExtrAns: Extracting answers from technical texts", IEEE Intelligence Systems, 18(4), pp. 12-17.
16. Lisa F Rau, "Conceptual Information Extraction from Financial News", IEEE TH0213-9/88, 1988.
17. Wilks, Yorick (1997), "Information extraction: Techniques and challenges", Lecture Notes in Artificial Intelligence, 1299, pp. 10-27.
18. Wu, Y., Liu, N.K.J. and Wang, K.Z. (1999), "An approach towards English automatic abstraction", *Journal of Computational Linguistics and Chinese Language Processing*, 4(1), pp. 85-102.
19. Zhou, L. and Zhang, D. (2002), "NLPIR: A framework for natural language processing in information retrieval", Journal of the American Society for Information Science and Technology, 54(2), pp. 115-123.
20. Zhou, L. and Liu, J. and Yu, S.W. (1998), "Automatic extraction of word pairs based on heuristics and statistical method", *International Journal on Computer Processing of Oriental Languages*, 11(4), pp. 339-351.

An Incremental FP-Growth Web Content Mining and Its Application in Preference Identification

Xiaoshu Hang[1], James N.K. Liu[2], Yu Ren[3], and Honghua Dai[1]

[1] School of Information Technology, Deakin University, Melbourne, Australia
[2] Department of Computing, Hong Kong Polytechnic University, Hong Kong
[3] Beijing Capital Science and Technology Group Co., Ltd, Beijing, China

Abstract. This paper presents a real application of Web-content mining using an incremental FP-Growth approach. We firstly restructure the semi-structured data retrieved from the web pages of Chinese car market to fit into the local database, and then employ an incremental algorithm to discover the association rules for the identification of car preference. To find more general regularities, a method of attribute-oriented induction is also utilized to find customer's consumption preferences. Experimental results show some interesting consumption preference patterns that may be beneficial for the government in making policy to encourage and guide car consumption.

1 Introduction

In the last decade, much attention has been paid to the studies on the theory and approaches of data mining, whereas only very few real applications of data mining can be seen from literatures. With the explosive growth of information sources available on the World Wide Web, it has become increasingly significant for users to acquire knowledge from WWW. Web mining can be generally classified into content mining, usage mining and web structure mining. Web content mining can be further classified into agent-based, including intelligent search agent and information filter agent, and database-based, including multilevel database and Web query system. Database-based approach focuses on techniques for organizing the semi-structured data into structured ones and saving them into database for later data mining [Cooley, R. *et al*, 1999].

Car consumption is viewed as a new driving power to the Chinese economic growth and paid high attention by Chinese government. Some car dealers distribute their car market information through Internet so that it is convenient for customers to acquire the updated information. Naturally the customers usually access the web pages of their favorite cars when they plan to purchase. Therefore, we assume that (1) the more a web page introducing a type of car is accessed, the higher this type of car is preferred by customers; (2) in general, a customer can manage car market information through Internet.

The semi-structured data retrieved from Web pages are distributed by Chinese largest car market in URL: http://www.cheshi.com.cn/ and mainly consist of the fields such as car producer, products type, product size, price, pricing date, the number of customers' visiting and other less valuable attributes. This web-based database in-

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 121–127, 2005.

cludes various kinds of cars, from lower price to very expensive ones. The number of customers accessed to a type of car ranges from less than one hundred to thousands, which shows different customers preference.

2 Web Data Acquisition

In this section, we discuss how to retrieve the interesting data embedded in web pages, such as HTML pages, XML pages or text files. Reconstruct the semi-structured data into more structured data is the first step of Web content-based mining. The popular method is to write a program called "Wrapper"[Kushmerick, N. *et al*, 1997, Adelberg, B., 1998] to extract data from web pages, and then store them in database. The web data acquisition includes: Input interesting URL and key words; fetching the Web pages; Trimming and normalizing the source file. The output is the wanted and normalized data blocks. The URL and keywords are viewed as input parameters of the preprocessing function which returns the wanted HTML or XML source files. Such raw data are then normalized to facilitate the next phase processing.

Discovering Interesting Region. A module called hook is designed to identify the interesting region from the normalized document. The hook algorithm is as follows:

```
Algorithm: Hook
  Input:    Keywords, a normalized document;
  Output:  The interesting region;
  Begin
     For I = 1 to m    // m is the number of keywords
        Calculating the occurrence number $n_i$ ,and  position $p_i = \{p_{i1}, p_{i2}, \ldots, p_{in}\}$ of
        ith keyword in the normalized document;
        If $n_i = 1$ then  return this table pair region;
        Else   find the minimum $n_k = \min(n_i)$.
            For j = 1 to $n_k$
              Calculating the jth keyword's table pair's positions tbegin, tend,
            which
              must satisfy  $tbegin < p_{kx} < tend$. Test all other keywords' position;
              If one of these keywords has no position within this region(tbegin,
            tend), continue next j, or  if all of them has at least one position in re-
            gion(tbegin, tend), return this region as the  interesting region.
         Endfor
      Endfor
End
```

Generating Structured Data. Inside the interesting data region, we are mainly concerned the three kinds of tags: "table", "tr" and "td". All other tags can be trimmed to obtain a skeleton of the data block. These tags reflect the hierarchy of data. As an example, a hierarchy can be generated as follows:

```
<table>
 <tr>
   <td currency </td> <td>  buy </td> <td> sale </td>
 </tr>
```

```
  <tr>
    <td> US$ </td> <td 826.47 </td> <td> 828.95 </td>
  </tr>
……
</table>
```

Then we pick up the data between <td> and </td> pair, store them in a pre-defined structured array and save them in local database for data mining.

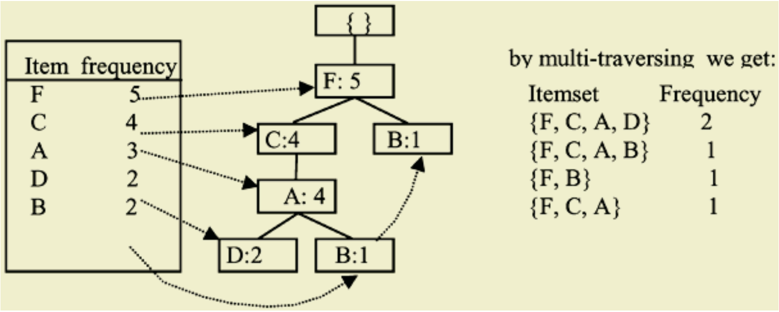
3 An Incremental FP-Growth Algorithm

FP-Growth algorithm is a highly efficient association rule mining algorithm. It finds association relationships among various items by compressing a large database into a small data structure and just scan database twice to build a FP-tree. The incremental association rule mining refers to the approach that updates the set of mined association rules with the new coming data flow.

The FP-Growth Based Incremental Approach. Let FT_0 be a FP-tree built by FP-Growth algorithm from data set D_0 at time t_0 and a new data flow ΔD are retrieved from web page in the time interval Δt, the problem is how to update FT_0 to obtain a new FP-tree FT_1 with data flow ΔD at time $t_1 = t_0+\Delta t$ so that FT_1 is completely the same as that built from data set $D_0+\Delta D$.

We have known that in FP-Growth algorithm, a header table is created for keeping each item's frequency in database. So the first step of the incremental FP-Growth algorithm is to update the header table with the new coming data set ΔD. This can be done just by scanning data set ΔD once and by resorting the items' frequency. Then the algorithm needs to scan data set ΔD again to create an initial FP-Tree FT' so that we have two FP-trees in memory. The FT' is much smaller than FT_0, because ΔD is much smaller. The next is what we called *FP-Contract and FP-Growth*. The basic idea is to grow FP-Tree FT' by contracting FT_0:

a) Multi-traversing FP-Tree FT_0 to generate all the item sets and their frequencies. An itemset's frequency is defined as the frequency of the terminal node in the itemset.
b) Contracting FT_0 by deleting the sub-trees in which the frequencies of all nodes are zero;
c) Growing FT-Tree FT' with the itemsets got from a);



Item	frequency
F	5
C	4
A	3
D	2
B	2

by multi-traversing we get:

Itemset	Frequency
{F, C, A, D}	2
{F, C, A, B}	1
{F, B}	1
{F, C, A}	1

Fig. 1. Multi-traversing a FP-Tree to get all the item sets

We get four item sets by traversing the above tree twice. Itemsets {F, C, A, D},{ F, C, A, B}and {F, B} are generated at the first traversing. Their frequencies are 2, 1 and 1, respectively. The tree becomes smaller by deleting the terminal nodes D and the two B, then it contains only three nodes F:1, C:1 and A:1 and an itemset {F, C, A} with frequency 1 is obtained by the second traversing.

Suppose we have obtained a new data set ΔD={{D, C, B},{D, A, C},{F, C, E}}from Web pages. So we construct a new FP-Tree FT′ with ΔD and the updated header table. (see Figure 2)

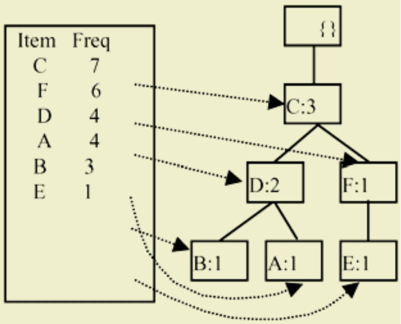


Fig. 2. FP-Tree FT′ built with ΔD

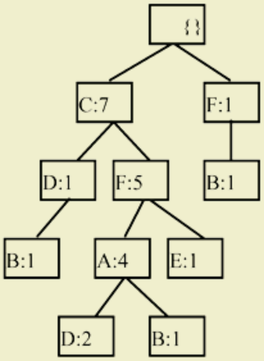


Fig. 3. FT′ after FP-Contraction and FP-Growth

FP-Tree FT′ continues to grow by inserting the itemsets gotten from FT_0 into it. Meanwhile the FP-Tree FT_0 is generally contracting and completely disappeared when the FT′ ends its growth. It can be proved that without considering the minimum support, that is, all the items in the dataset are inserted into the FT-Tree, the tree FT′ built by FP-Contracting and FP-Growing is completely the same as that built from the total dataset.

The *FP-Contraction and FP-Growth* algorithm for incrementally mining association rules can be described as follows:

Algorithm: *FP-Contraction and FP-Growth* for Incrementally Mining Association Rules

```
Input: a FP-Tree $FT_0$, a new retrieved data set $\Delta D$.
Output: a new FP-Tree FT′;
Begin
    For i= 1 to $|\Delta D|$   // $|\Delta D|$ represents the itemset number of $\Delta D$.
        For each $item_j$ in itemset $\Delta D_i$
            If $item_j$ occurs in the header list then its frequency +1;
        Endfor j
    Endfor i
    Delete the connection between the header list and the FP-Tree $FT_0$;
    Sorting the header table;
    Create a new FP-Tree FT′ from $|\Delta D|$;
    While ( $FT_0$ has child nodes )
```

```
        Depth-first traversing $FT_0$ to get a set of itemsets IS;
        For k=1 to $|IS|$  // for each itemset $IS_k$ in IS
        Rearrange the items in $IS_k$ according to the order of items in the header list;
        Insert each item in $IS_k$ into FP-Tree $FT'$ as the child of current node;
            Update the frequency of the nodes in $FT'$ with the frequency of $IS_k$ ;
          Connect  the new tree $FT'$ to the header list;
          Subminus the frequency of $IS_k$ from that of its each node in the path of $FT_0$;
            If a node with frequency 0 then delete it and its subtree;
        End for
    End while
    Generate association rules from FP-Tree $FT'$ ;
End
```

Evidence shows that the complexity of this algorithm is dominated by the complexity of FP-growth since the new coming data set is much smaller than the original database. We now consider mining in a dynamic data environment, the frequency of each item varies with the size of the total data set. So some items which are infrequent in the initial database may frequently occur in the incremental data set, while the other may never appear afterwards. To solve this problem, one method is to put all the items into the FP-Tree which is very expensive both in time and space, and therefore impractical.

4 Concept Generalization for Knowledge Abstraction

In this paper, three concept hierarchies (see Fig. 4) are designed in our mining task. The continuous attribute *price* is divided into 9 intervals and 5 higher level concepts are designed for concept induction. The attribute *model_name* is a category one and has a number of values which are classified into Chinese car and foreign car. The attribute *daily_access* is also continuous which is divided into six intervals and is further generalized three concepts.

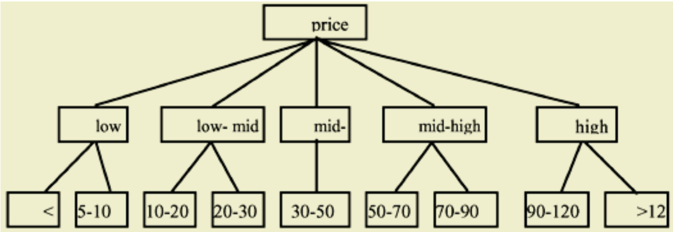


Fig. 4. Concept hierarchies for the attributes *Price, Car-name* and *Access*

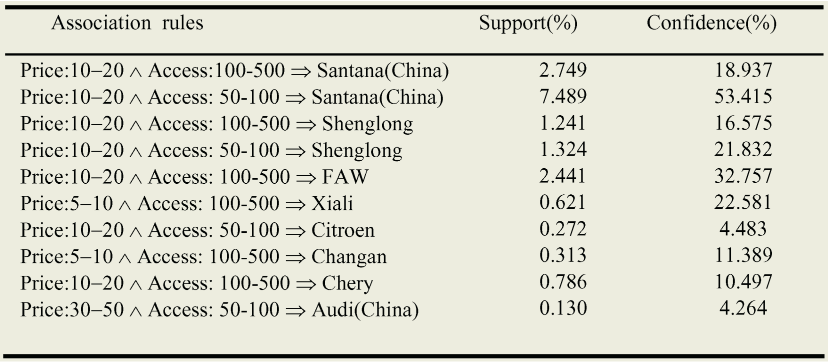
5 Application

In this section, we introduce a real-life application of our approach. The database retrieved by a wrapper from the web page of Chinese biggest car market http://www.

cheshi.com.cn, contains 11 fields among which the fields: *model_name, sub_ category, price, pricing_date* and *daily_access* are valuable to our mining task.

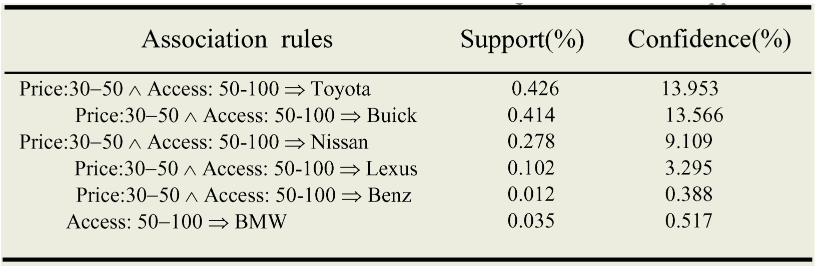
The association rules in Tables 1-3 are mined from the above database when it reaches the size of about 17500 instances. Table1 lists some rules that have strong association between attributes *price*, *daily_access* and *model_name*. They reflect the customers' higher preference to Chinese cars with price in the range of 100,000-200,000(RMB) than to those in other price ranges and *Satana* and *Shenglong* are the two major models that Chinese people prefer most. Table 2 shows that Chinese customers are most interested in foreign cars of moderate price ranging from 300,000 to 500,000(RMB). The association between *price* and *daily_access* is uncovered in Table 3 from which we can see that inexpensive cars are highly associated with frequent daily access whereas expensive cars have strong relation with less daily access. But the cars with price under 50,000(RMB) is an exception because they show no attraction to customers. Moderate price cars are moderately preferred by Chinese peoples.

Table 1. Preference to main models of Chinese cars with min_supp=0.1



Association rules	Support(%)	Confidence(%)
Price:10–20 ∧ Access:100-500 ⇒ Santana(China)	2.749	18.937
Price:10–20 ∧ Access: 50-100 ⇒ Santana(China)	7.489	53.415
Price:10–20 ∧ Access: 100-500 ⇒ Shenglong	1.241	16.575
Price:10–20 ∧ Access: 50-100 ⇒ Shenglong	1.324	21.832
Price:10–20 ∧ Access: 100-500 ⇒ FAW	2.441	32.757
Price:5–10 ∧ Access: 100-500 ⇒ Xiali	0.621	22.581
Price:10–20 ∧ Access: 50-100 ⇒ Citroen	0.272	4.483
Price:5–10 ∧ Access: 100-500 ⇒ Changan	0.313	11.389
Price:10–20 ∧ Access: 100-500 ⇒ Chery	0.786	10.497
Price:30–50 ∧ Access: 50-100 ⇒ Audi(China)	0.130	4.264

Table 2. Preferences to main models of foreign cars with min_supp=0.01



Association rules	Support(%)	Confidence(%)
Price:30–50 ∧ Access: 50-100 ⇒ Toyota	0.426	13.953
Price:30–50 ∧ Access: 50-100 ⇒ Buick	0.414	13.566
Price:30–50 ∧ Access: 50-100 ⇒ Nissan	0.278	9.109
Price:30–50 ∧ Access: 50-100 ⇒ Lexus	0.102	3.295
Price:30–50 ∧ Access: 50-100 ⇒ Benz	0.012	0.388
Access: 50–100 ⇒ BMW	0.035	0.517

6 Conclusion

To identify customer preference from web content data, we proposed an incremental association rule mining algorithm called FP-Contraction and FP-Growth. It rebuilds a new FP-Tree from the new coming data set and then iteratively breaks up the previous FP-Tree into itemsets and grows the new FP-Tree with these itemsets. Since the

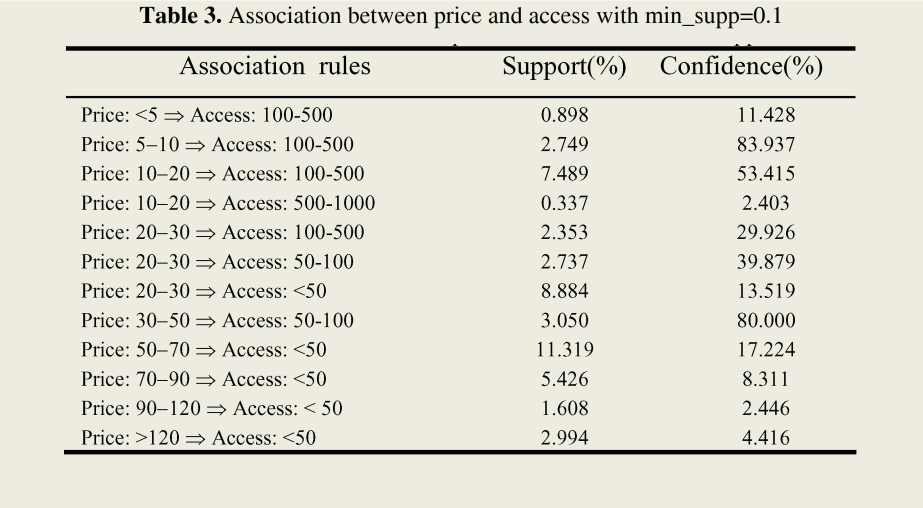


Table 3. Association between price and access with min_supp=0.1

Association rules	Support(%)	Confidence(%)
Price: <5 ⇒ Access: 100-500	0.898	11.428
Price: 5–10 ⇒ Access: 100-500	2.749	83.937
Price: 10–20 ⇒ Access: 100-500	7.489	53.415
Price: 10–20 ⇒ Access: 500-1000	0.337	2.403
Price: 20–30 ⇒ Access: 100-500	2.353	29.926
Price: 20–30 ⇒ Access: 50-100	2.737	39.879
Price: 20–30 ⇒ Access: <50	8.884	13.519
Price: 30–50 ⇒ Access: 50-100	3.050	80.000
Price: 50–70 ⇒ Access: <50	11.319	17.224
Price: 70–90 ⇒ Access: <50	5.426	8.311
Price: 90–120 ⇒ Access: < 50	1.608	2.446
Price: >120 ⇒ Access: <50	2.994	4.416

new coming data set is much smaller than that previously accumulated, it is relatively easier and faster to build the new FP-Tree. This new approach is designed to mining association rules in a dynamic data environment which receives new data flows from web pages progressively. A real-life application of our method is introduced which aims at discovering the car consumption preference. This application of web-content mining has discovered some interesting information of car preference. The system has been developed with VC++ in the environment of windows 2000 and has been put into practical use.

References

1. Cooley, R, Mobasher, B & Srivastava, J(1999). Web Mining: Information and Pattern Discovery on the World Wide Web,
   http://www-users.cs.umn.edu/~mobasher/webminer/survey/survey.html.
2. Kushmerick,N. Daniel, Weld, S. & Doorenbos,R.,(1997) *Wrapper induction for information extraction,* Proceeding of the 15th International Joint Conference on Artificial Intelligence. (pp. 729-737).
3. Kushmerick, N.(1997). Wrapper induction for information extraction. Ph.D. Dissertation, Dept. of Computer Science, Univ. of Washington, TR UW-CSE-97-11-04.
4. Adelberg,B.,(1998) NoDOSE - A tool for semi-automatically extracting structured and semistructured data from text documents, Proceedings of SIGMOD'98. (pp. 283-294).
5. Liu,L., Pu,C., & Han,W. *(2000)* Xwrap: An XML-enabled Wrapper Construction System for Web Information Sources , International Conference on Data Engineering, San Diego, CA.
6. J. Han, J. Pei, & Y. Yin, “Mining Frequent Patterns without Candidate Generation", SIGMOD Conference 2000: 1-12.
7. Webb, Geoffrey L., “Efficient search for association rules”. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 99--107, 2000.
8. Rakesh Agrawal & Ramakrishnan Srikant: “Fast Algorithms for Mining Association Rules in Large Databases”, VLDB 1994: 487-499.

Similarity Retrieval from Time-Series Tropical Cyclone Observations Using a Neural Weighting Generator for Forecasting Modeling

Bo Feng and James N.K. Liu

Department of Computing, the Hong Kong Polytechnic University
{csbfeng,csknliu}@comp.polyu.edu.hk

Abstract. Building a forecasting model for time-series data is a tough but very valuable research topic in recent years. High variation of time-series features must be considered appropriately for an accurate prediction. For weather forecasting, which is continuous, dynamic and chaotic, it's difficult to extract the most important information present in the knowledge base and determine the importance of each feature. In this paper, taking tropical cyclone (TC) as an example, we present an integrated similarity retrieval model to forecast the intensity of a tropical cyclone using neural network, which is adopted to generate a set of appropriate weights for various associated features of a tropical cyclone. A time adjustment function is used for time-series consideration. The experimental results show that this integrated approach can achieve a better performance.

1 Introduction

Enormous research efforts have been devoted into tropical cyclone (TC) forecasting in recent decades. Lee and Liu [1, 2] have presented an EGDLM system which automates the satellite interpretation process and provides an objective analysis of tropical cyclones. Their other research works [3, 4] have also exploited the chaotic features of neural oscillators, noting their exceptional object segmentation capabilities. In our previous works, we also obtained promising experimental results for matching two TC samples based on their significance-based shape points [5] and the distinctive spiral features among these points [6].

The last two decades have seen attempts to solve non-linear forecasting problems using AI technologies such as neural networks, fuzzy logic, genetic algorithms and expert systems. For instance, in [7], the authors allow a user to prescribe a solution based on the context of the current problem and those of selected past samples [8]. In [9, 10, 11], the authors proposed their approaches to allow numerical features to be converted into fuzzy terms and have greater flexibility in the retrieval of candidate cases. However, few of these research works ever tackle the intensity prediction of the tropical cyclone very well, and the critical obstacles to determine the importance of various features also exist. For TC prediction, it has its particular characteristics, which are continuous, data-intensive, multidimensional dynamic and chaotic. Besides, tropical cyclones have great relationship with time-series, which increase in complexity but haven't drawn too much attention from researchers. In this paper, we present a similarity retrieval model to predict TC intensities using a feed-forward neural weight generator, which is adopted to generate a set of appropriate weights for various associated features of a tropical cyclone. We also propose the time-series similarity ad-

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 128–134, 2005.

justment to measure the similarity of samples on consecutive observations of a tropical cyclone. To validate our idea, we have evaluated its performance against other retrieval methods. Experimental results show that our propose model has achieved a more precise output.

The rest of the paper is organized as follows. Section 2 presents the overview of the proposed model as well as the similarity measures and the neural weight generator. Experimental results and analysis are given in section 3. Section 4 concludes the paper and gives the future works.

2 System Overview

Figure 1 shows a forecasting model to predict the TC intensities. The model uses a neural weight generator to estimate the significance of every feature of a tropical cyclone from observation. The time series adjustment is used to improve the similarity calculation among the TC samples based on their observation sequence, which will be discussed in latter parts.

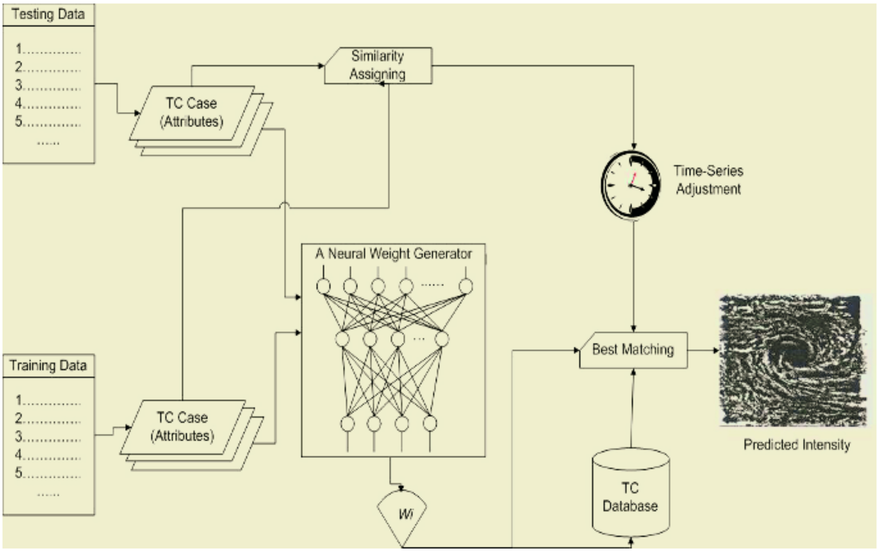


Fig. 1. Overview of the proposed prediction model

2.1 Similarity Measure

In Figure 1, for every observation sample, either from the training set or the testing set, it is fed into a neural weight generator to get a set of most appropriate weights for the attributes. There are several attributes representing the features of a TC sample, such as Name, Time, Intensity, Position, Speed, and MSLP and so on. Based on these time-series, positional and numerical attributes, we design different similarity measure functions for them respectively. Integrity of the weights and the similarity measures is used to carry out the prediction.

2.1.1 Measuring Function

An observation can then be defined as an attribute vector *C = [TN, TT, I, P, S, Sp]*, where each symbol represents a corresponding attribute for a TC observation. We only consider numerical variables such as Position, SMLP and Speed for similarity measures. The similarity of two observations of attributes SMLP and Speed can be determined by their distance measure, as shown in equation (1).

$$D_{SMLP/Speed} = d_{ij} = \frac{|A_i - A_j|}{\max(A) - \min(A)} \tag{1}$$

where A_i and A_j are values of the attribute *A* in different observations, and *max(A)* and *min(A)* are the maximum and minimum respectively in all samples of the attribute *A*. For variable Position, as it contains two parameters representing its orientations, the distance can then be measured as:

$$D_{position} = d_{ij} = \frac{|P_i - P_j|^2}{(\max X)^2 + (\max Y)^2} \tag{2}$$

where P_i and P_j are the Position vectors *[latitude, longitude]* in different observations, and *maxX* and *maxY* denote the most right-bottom point in the Latitude-Longitude coordinate system covered in our database. Therefore, we get the similarity functions for these three attributes as:

$$Sim(S_i, S_j) = 3 - D_{SMLP} - D_{Speed} - D_{position} \tag{3}$$

2.1.2 Neural Weight Generator

After we have determined the similarity measure functions, we calculate the weight of every attribute of each TC observation instance. The weight is used to determine the importance of each attribute contributing to the similarity of two observation instances. In this work, neural network technology is adopted to calculate the weight for each attribute, by means of predicting the attribute Intensity for the observation samples in the database. First of all, the values of these attributes need to be normalized and changed into binary forms. We use equal width bin approach to normalize each attribute, and the bin width is calculated using equation 4:

$$width = \frac{\max(att) - \min(att)}{bin} \tag{4}$$

where *max(att)* and *min(att)* are corresponding to the maximum and minimum value of that attribute among all observations respectively, and *bin* is the number of bins to be divided into. Taking attribute MSLP with value 100 as an example, it is supposed to be located at the 10th bin, out of totally 10 bins of MSLP. So, value 100 can be transformed into 0000000001, where the binary bit 1 represents its bin number. We classify tropical cyclones into four categories (tropical depression, tropical storm, severe tropical storm and typhoon) and use a four-bit binary number to represent it like 0100, denoting the given sample that belongs to the second category.

After the input pattern has been normalized and well-represented, it is fed into a feed-forward neural network for training and validation with sigmoid activation function. Assuming that instances of the input observations are $A_k=(a_k^1, a_k^2, \ldots, a_k^n)$ ($k =$

1,2,..., m), where k is the index for the input observations, and *n* is the number of the input unit. The hidden layer unit *j* computes its activation values as below:

$$o_j = f(\sum_{i=1}^{n} a_k{}^i w_{ij} - \beta_j) \tag{5}$$

where β_j is the bias of the j^{th} unit of the hidden layer, $a_k{}^i$ is the input instance, w_{ij} is the feed-forward weight connecting hidden unit *j* to input unit *i*, and f is the sigmoid function. w_{ij} is randomly distributed in [0.0, 1.0]. When the training progress has completed for a given observation sample, a set of weights w_{ij} can be achieved for every input node *i* connecting to every hidden unit *j*. These weights can be regarded as the importance degree of every input node *i,* which is the similarity attribute of the observation samples, contributing to determining the network output, intensity classification. After that the similarity of two attribute vectors C_i and C_j can be modified as:

$$Sim(C_i, C_j) = \sum_{k=1}^{N} [(\frac{w_{ik} + w_{jk}}{2}) * Sim(S_{ik}, S_{jk})] / \sum_{k=1}^{N} (w_{ik} + w_{jk}) \tag{6}$$

where S_{ik} and S_{jk} are values of k^{th} attribute for two observations respectively.

2.1.3 Time-Series Measure Function

A tropical cyclone will be recorded with more than one observation during its life for the analysts' prompt and precise references. The more observations for a tropical cyclone are recorded, the more precise prediction results will be achieved, and the earlier we can take action to avoid the disaster for loss of human lives and properties. Besides, empirically the form of the tropical cyclones is sensitive to time in any region. For example, in Hong Kong, there are more tropical cyclones with stronger intensities from June to August every year. Therefore, the attribute Time will have its particular impact to the similarity measure during the retrieval. In this subsection, a time function E(-t) is introduced because of its depression properties: E(0) = 1 and E(- ∞) = 0. Details of this time function are given as below:

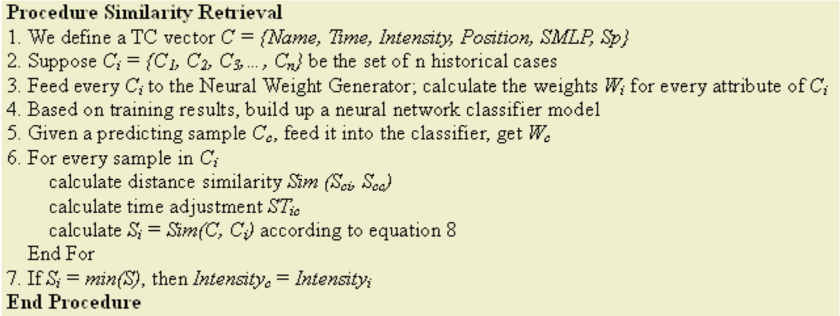
$$ST_{ij} = \exp\left[- | (\frac{Month_i}{12} + \frac{Day_i}{30} * \alpha) - (\frac{Month_j}{12} + \frac{Day_j}{30} * \alpha) | \right] \tag{7}$$

where *Month*$_i$ *(Month*$_j$*) and Day*$_i$ *(Day*$_j$*)* are the recording times for observations *i* and *j*. To show the dominance of variable Month over variable Date, a parameter α on Date is adopted to lower its influence to ST_{ij}, where α belongs to [0.0, 1.0]. One of the advantages is that we avoid simply modeling the periodic effect in monotonously increasing or decreasing mode, instead of which a step-wise mode is used. Another advantage is that as *max(ST) = 1* and *min(ST) = 0*, then the result of ST_{ij} will not change rapidly even though two recorded times are at a longer time interval of each other, which makes it more feasible to be integrated into the previous similarity functions. Consequently, that the similarity of two attribute vectors C_i and C_j can be further modified in equation (8).

$$Sim(C_i, C_j) = \sum_{k=1}^{N} [(\frac{w_{ik} + w_{jk}}{2}) * (ST_{ij} + Sim(S_{ik}, S_{jk}))] / \sum_{k=1}^{N} (w_{ik} + w_{jk}) \tag{8}$$

3 Experimental Results and Analysis

In order to evaluate the usefulness of our proposed competitive neural network classifier, we carry out a set of experiments to test whether our approach achieves an acceptable prediction precision, and whether there is a superiority comparing to other existing forecast models. We have collected 6,687 observation samples of 324 tropical cyclones passing through Hong Kong in the ten years from 1994 to 2003. These samples are time-series distributed and recorded every 6 hours during their lives. A computer with 2.26 GHz Intel Pentium CPU and 512M RAM memory is used for the simulations. As we can not afford to take all these samples in our experiments, to ensure the representative, we randomly select 200 observation samples as the training set for the neural weight generator, and select randomly another 30 samples as the testing set. Figure 2 gives a detailed similarity retrieval procedure used in our experiments.



```
Procedure Similarity Retrieval
1. We define a TC vector $C = \{Name, Time, Intensity, Position, SMLP, Sp\}$
2. Suppose $C_i = \{C_1, C_2, C_3, ..., C_n\}$ be the set of n historical cases
3. Feed every $C_i$ to the Neural Weight Generator; calculate the weights $W_i$ for every attribute of $C_i$
4. Based on training results, build up a neural network classifier model
5. Given a predicting sample $C_c$, feed it into the classifier, get $W_c$
6. For every sample in $C_i$
       calculate distance similarity $Sim\ (S_{ci}, S_{cc})$
       calculate time adjustment $ST_{ic}$
       calculate $S_i = Sim(C, C_i)$ according to equation 8
   End For
7. If $S_i = min(S)$, then $Intensity_c = Intensity_i$
End Procedure
```

Fig. 2. Steps to predict the intensity of current sample

Before we feed the similarity attributes into the neural weight generator, we normalize them into binary representation as mentioned in section 2.1.2. Table 1 shows the normalization results and the number of units for every network layer. In total we need 120 bits to represent the input values of three variables; we put 120 neurons in the input layer. Empirically the number of the hidden units needed is equal to one tenth of the number of input units, so 12 nodes are put into the hidden layer. In the last layer, depending on the predicted intensity category, we use 4 units in this layer to show the classified results from the neural network classifier.

Table 1. Input attributes normalization and initial number of layer units for neural weight generator

Input Pattern				Initial number of		
				Input units	**Hidden units**	**Output units**
Attribute	**No.**	**Example**	**Representation**	*120*	*12*	*4*
Position	*Long (10) x Lat (10) = 100*	*N_{ij}^{th} block, i=20,j=20*	*100…0 (ninety-nine 0s)*			
MSLP	*Bin(10) x 1 = 10*	*109*	*100…0 (nine 0s)*			
Speed	*Bin(10) x 1 = 10*	*136*	*100…0 (nine 0s)*			

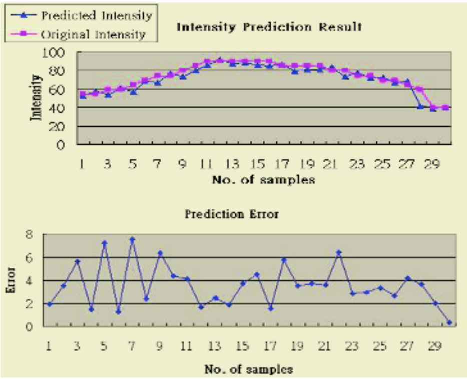


Fig. 3. Intensity prediction results and errors

In Figure 3 we show the predicted results and errors for 30 selected TC observation samples. We can clearly see that the predicted intensities are very close to the original intensities in the database. The largest difference between the predicted intensity and the original one is 8, which is considered acceptable.

To justify the performance of our propose model, another two existing TC forecasting models developed in our previous works are used for comparison: 1) Weighted Hausdorff Distance Measure [5], and 2) Spiral Feature Comparison [6]. The totally 30 observation samples selected from the database are used for comparison, which is calculated using following equation:

$$P = \frac{|I_p - I_r|}{I_r} \tag{9}$$

where I_p is the predicted intensity and I_r is the actual intensity for that observation in the database. As shown in Figure 4, our proposed retrieval model has the highest predict precision among all three approaches at most of samples. For the proposed retrieval model solely, the prediction precision is above 80%, which is considered that the proposed method is feasible.

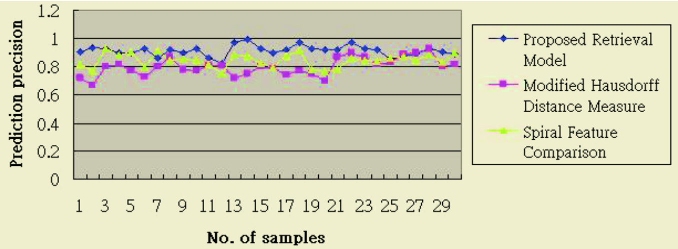


Fig. 4. Predicted precision of three approaches

4 Conclusion and Future Works

We have presented a TC intensity forecast model based on similarity retrieval from non-linear tropical cyclone observations. Neural network is adopted to generate a set of appropriate weights for various feature variables of a tropical cyclone instance. We also propose the time-series adjustment function to measure the particular impact of the variable Time. Ten years of data comprising 6,687 observation samples of 324 tropical cyclones passing through Hong Kong is used for the experiments. Results show that our proposed approach has achieved a more precise output with acceptable computational cost. Future research efforts will be directed towards improving the similarity function for TC attributes, and enhancing the measure for the time series impact among different observation samples which are imprecise in nature. Therefore, investigation on the use of fuzzy theory and case base adaptation to counteract exceptional TC behavior can be useful in the course of accurate forecast.

Acknowledgement

The authors would like to acknowledge partial support through the research grant CERG B-Q515 and departmental research grant H-ZJ87 of The Hong Kong Polytechnic University.

References

1. Lee, R.S.T. and Liu, J.N.K.: An automatic tropical cyclone pattern recognition and classification system using Composite Neural Oscillatory-based EGDLM, Journal of Fuzzy Systems, Vol. 4, No. 1 (2002) 616-625
2. Lee, R. S. T. and Liu, J.N.K.: An Elastic Contour Matching Model for Tropical Cyclone Pattern Recognition, IEEE Transaction on Systems, Man, and Cybernetics, Vol. 31, Part B, No. 3 (2001) 413-417
3. Liu, J.N,K, Raymond W.M.Kwong, Meng Wang, Danny K.Y. Sin and Lakhmi C. Jain.: An Integrated Approach for the Prediction of Tropical Cyclone and Weather Forecast, WSEAS Transactions on Systems, Vol. 2, Issue 3 (2003) 716-723
4. Lee, Raymond S.T. and Liu, J.N.K.: iJADE WeatherMAN – A weather forecasting system using intelligent multi-agent based fuzzy-neuro network, IEEE Transaction on Systems, Man, and Cybernetics, Vol. 34, Part C, No. 3 (2004) 369-377
5. Feng B. and Liu, J.N.K.: Semo-Mamo, A 3-phase module to compare tropical cyclone satellite images using a modified Hausdorff distance, In the Proceeding of the International Conference on Machine Learning and Cybernetics (2004)3808-3813
6. Liu, J.N.K. and Feng B.: A Novel Comparison Approach for Tropical Cyclone Satellite Images using Angle Features, to appear in the Proceeding of 18th International Florida Artificial Intelligence Research Society Conference (2005)
7. Aguirre, J.L., Montano, O., Sanchez-Castellanos and J.M.: Knowledge flow leveraged through case-based and data mining agents in a just in time information and knowledge system, In the Proceeding of International IEEE Conference on Intelligent Systems, Vol.1 (2004) 200 - 205
8. Kolodner, J.: Case-Based Reasoning, Morgan Kaufmann, California, USA, 1993.
9. Louis, S.J. and McDonnell, J.: Learning with case-injected genetic algorithms. IEEE Transactions on Evolutionary Computation, Vol.8, Issue.4 (2004) 316 - 328
10. Pal, S.K. and Pabitra Mitra: Case generation using rough sets with fuzzy representation. IEEE Transactions on Knowledge and Data Engineering, Vol.16 (2004) 293 - 300
11. De Calmes, M., Dubois, D., Hullermeier, E., Prade, H. and Sedes, F.: Case-based querying and prediction: a fuzzy set approach. In the Proceeding of IEEE International Conference on Fuzzy Systems, Vol.1 (2002) 735 - 739
12. Hong Kong Observatory, HKO: http://www.hko.gov.hk/contente.htm

Web Access Path Prediction Using Fuzzy Case Based Reasoning

Simon C.K. Shiu* and Cody K.P. Wong

Department of Computing, Hong Kong Polytechnic University, Hong Kong, China
csckshiu@comp.polyu.edu.hk

Abstract. In this paper, a fuzzy case based reasoning approach to Web access path prediction is developed and tested. It is based on the assumption that new user's access patterns can be predicted by referencing to the behaviors of similar Web users in the past. This method has three phases. Firstly, a Web case base is constructed from the Web log data. This includes the pre-processing and cleaning of the Web log data so that a suitable format is developed. Secondly, contextual information is extracted from the Web pages, and this information is used to develop a similarity measurement between Web pages. This information is added to the Web case base. Finally, fuzzy association rule mining is used to discover the relationship between the browsing behavior (user navigations) and the Web contents using the Web case base. A set of predictive cases from the Web case base is then selected for the access path prediction. From the experimental evaluation, our approach has demonstrated better prediction accuracy than the existing approaches.

1 Introduction

Web mining can be broadly defined as the automatic discovery and analysis of useful information from the World Wide Web [1]. In this paper, we adopt a Fuzzy Case based Reasoning (FCBR) approach [7] to develop our purposed Web access path prediction system. Case based Reasoning means reasoning from prior examples. It involves retaining a memory of previous problems and their solutions and, by referencing this knowledge, new problems are compared, and the previous successful solutions are adapted (i.e., modified) and applied to the new problem situation.

In order to integrate different sources of data to a common platform and format, features are extracted from data and then transformed to a unique and measurable form of representation. Apart from the mechanism of data integration, the similarity measure on the transformed data model becomes another critical issue. Different types of models have been constructed by focusing on the different characteristics of the data, for example, the feature characteristics, the inter- and intra-data structural characteristics, and the domain characteristics. Therefore, in applications that require a fusion of different sources and format of data, a careful integration mechanism and a revised similarity measure will be required. Furthermore, fuzzy set theory would be appropriate if the data sets are incomplete, uncertain or having overlap boundaries. Thus, the use of FCBR becomes apparent.

* This research is supported by the Hong Kong Polytechnic University research grant A-PD55

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 135–140, 2005.

2 Web Case Base Construction

Our approach to construct a Web case base from the Web usage data, i.e., the Web server log is as follows. A case in the Web case base consisted of two set of information, the information describing the problem and the information describing the solution. We could define the problem part as the set of pages that a user has accessed for the purpose of navigation for linking to the desire data. The solution part could be defined as the set of pages that actually contain the desired data. The access pattern of a certain type of user can be reflected by the length of the user session, and the future access path of a user is not simply related to the last visited URL of the user, but a series of accesses. Here, with reference to the case definition proposed by Yang [10], we define the length of a user transaction as the total number of page files that have been accessed within the transaction. They are discretized into N partitions by a 1-D linear clustering method based on the number of page files accessed per transaction.

3 Similarity Measurement of Web Content Information

Our approach of calculating similarity among web pages is based on the integration of different sources of Web content data to construct a knowledge base that can be used as a kind of support to the Web case base that is constructed from the Web log in Section two. This knowledge base supplies the Web content information in terms of a set of inter-page similarities for each possible pair of linked pages and this information is used for the mining the fuzzy association rules afterwards. This technique of measuring similarity among web pages is explained next.

3.1 Term Frequency Measurement

In the information retrieval literature, TF-IDF is the most commonly used and simplest measurement method [12] for feature extraction from documents. It is useful to use this t-dimensional term vector to represent a Web page. Absence of a term is indicated by a "0" and presence of a term is indicated by a positive number know as the weight. The normalized weighting functions for w_{ik} is:

$$w_{ij} = \frac{tf_{ij} \bullet idf_j}{\sqrt{\sum_{j=1}^{t} (tf_{ij})^2 \bullet (idf_j)^2}}$$

where w_{ij} is the weight of term j in Web page i, tf_{ij} is the term frequency of term j in Web page i. idf_j is the Inverse Web page Frequency of the term j in Web page i is computed by:

$$idf_j = \log\left(\frac{N}{n_j}\right)$$

where N is the number of Web pages in the case base, and n_j is the number of Web pages in the case base that containing the term t_j. After computing all the term weights, they will be used as the universe of discourse to formulate the linguistic variable "term weight importance". After this calculation, each web page is represented by a corresponding term vector. The hybrid similarity calculation using both the Euclidean and hierarchical distances is developed as follows:

$$SimI(p_1, p_2) = \frac{\alpha\ SimD(p_1, p_2) + \beta\ SimH(p_1, p_2)}{\alpha + \beta}$$

where SimI(p1, p2) is the integrated similarity between two web pages, p1 and p2.

3.2 Mapping to Concept Hierarchy

In general, concept maps are sorted graphs that are visually represented as nodes and edges with (or without) weights.

3.2.1 Construction of Concept Hierarchy from ODP Tree

To obtain a better mapping from the keyword-pairs to the concept hierarchy, the hierarchy should have the following characteristics if possible.

ODP was initially named "Gnuhoo", and it was established in early 1998 by Skrenta and Truel. It is maintained by many voluntary editors who have been carefully selected. The Open Directory Project has become a leading human-compiled open source Web directory. In year 2000, the ODP boasts more than 22,000 editors, with more than 4 million Web sites in over 400,000 categories. ODP data is also used by popular Web search sites, such as Alta Vista, AOL, HotBot, Lycos, and Netscape, as well as dozens of smaller search sites [15].

3.2.2 Adaptation of Concept Hierarchy Using WordNet

WordNet was developed by the Cognitive Science Laboratory at Princeton University under the direction of Professor George A. Miller. It is an on-line lexical reference system, and its design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept.

3.2.3 Mapping Keyword-Pairs to Concept Hierarchy

The procedures to map the extracted keyword-pairs to a concept hierarchy are summarized by the following pseudo-codes:

```
For each Web page I in Web site;
  For each keyword-pair J in Web page I;
    Repeat
    Searching on the concept hierarchy {
    If the keyword of the node match with keyword-pair IJ
    Then assign the keyword-pair with the position of the node
      Else if it is not the leaf node
      Then move to the next node;
      End if;
      }End repeat;
    Until a node is matched or a new node is added;
  End for J;
End for I;
```

Each keyword-pair's position in the concept hierarchy is identified by matching them to the concept hierarchy, if no match is found, then a new node is created.

3.2.4 Hierarchical-Based Similarity Measure

In CBR research, hierarchical-based similarity measures have been developed on cases which could be organized more effectively as a hierarchy [16][17]. Two major approaches are used to calculate the hierarchic similarity: one is based on which level of nodes the case was found matched, and the other approach is based on the total number of nodes connected to the matched case.

3.3 Complement Web Case Base by Content Similarity

In the previous Section two we have constructed the Web case base from the Web usage log and using the similarity measures defined above we can determine the content similarity between Web pages.

After the case base is constructed, our next step is to select predictive cases, among all the cases, for use in the actual access path prediction. This is discussed in the following section.

4 Mining Fuzzy Association Rule for Predictive Cases Selection

In the previous section, we give a hybrid similarity measure that operated using the keyword-pairs, and the computed inter-page similarity is used to complement the Web case base. In order to discover the fuzzy association rules from the Web case base, numeric similarity is fuzzified into five linguistic terms: Highly Similar, Quite Similar, Medium, Not So Similar, Not Similar, and these are represented by HS, QS, M, NSS, NS.

We define the following form for a fuzzy association rule:

IF $X = \{x_1, x_2, \ldots, x_n\}$ is $A = \{f_1, f_2, \ldots, f_n\}$
THEN $Y = \{y_1, y_2, \ldots, y_n\}$ is $B = \{g_1, g_2, \ldots, g_n\}$

where X is the sequence of URLs accessed and A is the associated fuzzy set, and Y is the sequence of URLs predicted and B is the associated fuzzy set.

X is the problem part of the case u_i in C_s and Y is the corresponding solution part of the case u_i. We want to find the fuzzy association rule in the form of "IF the inter-page similarity of X is A THEN the inter-page similarity of Y is B".

A fuzzy support value reflects not only the number of records supporting the itemset, but also their degree of support. We adopt the formula proposed by Gyenesei [23] to calculate the fuzzy support value of the itemsets of <X, A>.

We use the following example to illustrate the computation of the fuzzy support value. Let X ={URL1, URL2} and A ={Highly Similar, Medium}.

$$FS_{<X,A>} = \frac{\sum_{u_i \in C_s} \prod_{x_j \in X, f_j \in A} u_i : (x_j, f_j)}{|C_s|}$$

where $|C_s|$ is the cardinality of the case set.

Using the cases in Table 1, the fuzzy support value of the rule "IF the inter-page similarity of URL1 and URL2 is Highly Similar THEN the inter-page similarity of URL2 and URL3 is Medium" is calculated as follows:

$$FS<X,A>=\frac{0.5*0.8+0.6*0.6+0.4*0.8+0.7*0.2+0.5*0.6+0.2*0.4+0.9*0.1}{7}$$

$$= 0.241$$

After the fuzzy association rule mining, we obtained a set of fuzzy association rules. Next, we use the following steps to choose predictive cases from the Web case base.

INPUT: A set of fuzzy association rule, a set of fuzzy membership function, a set of inter-page similarity values and the site structure and a predefined minimum support value λ, and confidence value γ.

OUTPUT: A set of predictive cases for Web access path prediction.

5 Conclusion and Future Work

In this research, a case based reasoning approach for Web access pattern prediction is developed and tested. In developing of this framework, some new techniques and ideas are developed and implemented. Firstly, we proposed a set of techniques to construct a case base from the Web log data. This case base represents the Web usage data in terms of cases. In order to capture the content information of Web pages, we have explained the procedures and techniques to extract content information from the Web pages by transforming and mapping the Web content to a common representation. To measure the information extracted, qualitative (i.e., hierarchical information) measures is proposed and this measure is used to find the similarity between Web pages. In order to discover the relation between the navigation and the content browsed, a new fuzzy association rule mining algorithm is designed. It is used to discover the useful patterns in the Web log data based on the relation between browsing behavior and content. Procedures to select the predictive cases from the case base using the fuzzy association rules are described and the cases selected are used for access path prediction. The proposed algorithms and techniques are evaluated by experiments and results are given.

References

1. Cooley, R., Mobasher, B. and Srivastava, J.: Web Mining: Information and Pattern Discovery on the World Wide Web (A Survey Paper), in Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), (1997)
2. Madria, S.K., Bhowmick, S.S., Ng, W.K., Lim, E.P.: Research issues in Web data mining, in Proceedings of Data Warehousing and Knowledge Discovery, (1999) 303-312
3. J. Borges and M. Levene. Mining association rules in hypertext data base, in Proceeding of Fourth International Conference of Knowledge Discovery and Data Mining (KDD-98), August 27-31, New York City, USA, (1998)
4. J. Borges, and M. Levene. Data Mining of User Navigation Patterns, in Web Usage Analysis and User Profiling, pp. 92-111. Published by Springer-Verlag as Lecture Notes in Computer Science, Vol. 1836, (1999)

5. Z. Su, Q. Yang, and H. J. Zhang. A Prediction System for Multimedia Pre-fetching, in Proceedings of the eighth ACM Multimedia Conference (ACM Multimedia 00). Los Angeles, California, (2000)
6. Z. Su, Q. Yang, Y Lu, and H. J. Zhang. WhatNext: A Prediction System for Web Requests using N-gram Sequence Models, in Proceeding of the First International Conference on Web Information Systems Engineering (WISE'00), Hong Kong, (2000)
7. S. K. Pal, S.C.K. Shiu, Foundations of Soft Case-Based Reasoning, John Wiley, Hoboken, New Jersey (2004)
8. M. Doyle, M. A. Ferrario, C. Hayes, P. Cunningham, and B. Smyth, CBR Net:- Smart Technology over a Network, TCD-CS-1998-07 report, University College Dublin (1998)
9. A. Joshi, and R. Krishnapuram. Robust Fuzzy Clustering Methods to Support Web Mining, in Proceedings of ACM SIGMOD Workshop on Data Mining and Knowledge Discovery,(1998)
10. Q. Yang, H. H. Zhang, and H. Zhang. Taylor Series Prediction: A Cache Replacement Policy Based on Second-Order Trend Analysis, in Proceeding of the HICSS, (2001)
11. K. Racine, and Q. Yang, Maintaining Unstructured Case bases, AAAI Technical Report – Verification and Validation Workshop, (1996)
12. G. Salton, and C. Buckley, Term-weighting approaches in automatic text retrieval, in Information Processing & Management, Vol. 24, No.5, (1988) 513-523
13. W. Wang, W. Meng, and C. Yu, Concept Hierarchy Based Text Database Categorization in Metasearch Engine Environment, in Proceedings of the First International Conference on Web Information Systems Engineering (WISE'2000), Hong Kong, (2000), 283-290
14. G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, Introduction to WordNet: An on-line lexical database, International Journal of Lexicography, 3(4), (1990) 235-312
15. C. Sherman, and C. Human, Do it Better: Inside the Open Directory Project, ONLINE, (2000)
16. X. Hu, Conceptual Clustering and Concept Hierarchies in Knowledge Discovery, Thesis for the degree of Master of Science in the School of Computing Science, (1993)
17. E. Armengol, and Enric Plaza, Similarity Assessment for Relational CBR, Processing ICCBR 2001: (2001) 44-58.
18. B. R. Kosanovic, Temporal fuzzy sets. Retrieved March 21, 2002 from the World Wide Web: http://www.neuronet.pitt.edu /~bogdan/reserch/fuzzy/f/f.html
19. J. C. Bezdek, R. Erlich, and W. Full, The fuzzy c-mean clustering algorithm, Computers and Geoscience, Volume 10, No. 2-3, (1984) 191-203
20. H. J. Zimmerman, Fuzzy set Theory and its applications, Kluwer Academic Publishers, Massachusetts, USA, 5th Edition, (1992)

Multiple Classifier System with Feature Grouping for Intrusion Detection: Mutual Information Approach

Aki P.F. Chan, Wing W.Y. Ng, Daniel S. Yeung, and Eric C.C. Tsang

The Hong Kong Polytechnic University, Department of Computing, Hong Kong, China
{csaki,cswyng,csdaniel,csetsang}@comp.polyu.edu.hk

Abstract. The information security of computer networks has become a serious global issue and also a hot topic in computer networking researches. Many approaches have been proposed to tackle these problems, especially the denial of service (DoS) attacks. The Multiple Classifier System (MCS) is one of the approaches that have been adopted in the detection of DoS attacks recently. For a MCS to perform better than a single classifier, it is crucial for the base classifiers which embedded in the MCS to be diverse. Both resampling, e.g. bagging, and feature grouping could promote diversity of base classifiers. In this paper, we propose an approach to select the reduced feature group for each of the base classifiers in the MCS based on the mutual information between the features and class labels. The base classifiers being combined using the weighted sum is examined in this paper. Different feature grouping methods are evaluated theoretically and experimentally.

1 Introduction

Due to the tremendous increasing connectivity and accessibility to the Internet in recent years, intrusions and crimes related to computer network systems have become a serious global issue. According to the CSI/FBI survey in 2003, the financial losses due to network intrusions in the United States alone have increased greatly since 2002 and were exceeded US$450 Millions. The DoS is particularly interesting because of its big impact to those e-commerce systems or critical systems, which results in incalculable losses, trust and loyalty reduction for the e-commerce companies. The conventional approaches for intrusion detection could be divided into two categories: misuse and anomaly detection. Both of them possess different limitations and details can be found in [17,18].

The MCS is a young and active research area which gives better performance than a single classifier [1,2]. Many researchers found that just selecting a single classifier that performs well may not be the optimal choice. It may lose potentially valuable information contained in other less accurate classifiers. Thus the MCS approach is suggested as a solution which merges several less accurate classifiers. For a MCS to achieve higher accuracy than a single classifier, it is crucial that the base classifiers are sufficiently diverse [5,7,19]. Two methods could be used to promote the diversity which are resampling and feature grouping. The resampling approaches select different training samples for each base classifier and the feature grouping approaches select different groups of features for training each base classifier. Bagging and

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 141–148, 2005.

Boosting are the resampling techniques being widely applied in MCS training. However, feature grouping is still a relatively new topic in a MCS. One may notice that feature grouping is sometimes referred to the feature selection of MCS. However, we may distinguish the feature grouping as a selection method for feature group of individual base classifier in the MCS. In contrast, feature selection tries to remove particular features from the entire dataset and one does not need to collect this feature in future data collection [21]. Certainly, feature selection could be a side-product of the feature grouping if all feature groups do not involve a subset of features. In fact, a MCS trained using feature grouping could yield a better diversity than a MCS trained using resampling. In feature grouping, every base classifier tries to solve the pattern classification problem using different group of features and thus their solution and performance could be very different and diverse [6,7]. The major research problem in feature grouping is to find the best selection criterion for selecting the groups of features for each base classifier while maintaining the classification accuracy for unseen samples.

In DoS detection problem, one is required to maximize the classification accuracy and minimize the false alarm rate. So, a novel feature grouping method using mutual information for building a MCS is proposed in this paper to deal with the DoS problem.

A brief survey to current pattern recognition approaches in intrusion detection and the methods of selecting features for the base classifiers are given in Section 2. Section 3 describes the methodology of the proposed method for selecting feature groups for the base classifiers in the MCS and the experimental results and empirical comparisons are discussed in Section 4. Finally, Section 5 gives the conclusion of this work.

2 A Brief Survey on Related Works

2.1 Pattern Recognition Approaches for ID

The DoS detection problem is formulated as a pattern classification problem of classifying a sample to be a DoS attack or a normal sample. The sample in the DoS detection problem is composed by the current network information and the packet passing through the DoS detection system which has been installed in the network system. Neural network [20, 21], support vector machines (SVM) and multiple classifier systems (MCS) [3,4] have been applied in the DoS detection due to its ability in realizing arbitrary continuous mappings between the inputs and outputs.

2.1.1 Single Classifier Approach

Many researchers have applied neural network in intrusion detection [20, 21]. A neural network consists of a set of sensory units that merging together to implement the complex mapping function. A neural network learns and generalizes the DoS detecion from historical data, i.e. training dataset. SVM is another approach which maps the input feature values of samples into a higher dimensional feature space where the attacking and normal samples are linearly separable and solves the DoS detection problem by finding a decision hyperplane in the feature space.

2.1.2 Multiple Classifier System Approach

The MCS is proposed to solve DoS detection problems by the inspiration of the human experts trying to design attack signatures by combining different attack characteristics in order to attain high attack detection rates and low false alarm rates [3, 4]. It is also shown to have better performance than a single classifier [1, 2]. Feature grouping further improve the performance of a MCS by reducing the 'curse of dimensionality' in the samples and the correlation among the base classifiers by training them on different groups of features [6].

2.2 Feature Grouping Approaches of MCS

2.2.1 Random Approach

A group of features is randomly selected for each of the base classifiers in the MCS and the feature group size is chosen by an ad-hoc method [8]. Each base classifier is then trained on the randomly selected group of features and their results are combined by a fusion rule. This approach is realized as the Attribute Bagging [9] for feature grouping in a MCS. Although random approach could promote the diversity of a MCS, the classification accuracy of the MCS has not been considered.

2.2.2 Heuristics Approach

The features are divided into different groups based on the domain expert knowledge about the characteristics of attack signatures. Content features and network related features are suggested for the division of feature groups. The network related features are further divided into intrinsic and traffic features [3]. Each base classifier is then trained on one of the feature groups. The deficiency of this approach is that the domain expert may ignore some hidden relation between input features and the class labels for an attack class and overestimate the importance of some features.

2.2.3 Input Decimation

Input decimation generates different subsets of features for each of the base classifiers in the MCS. The number of base classifiers is equal to the number of classes in the problem. The features are selected for a particular base classifier if they yield a high absolute correlation coefficient to the class of this base classifier. Thus each base classifier is trained on the group of features which carry strong discriminating information for a particular class [6]. However correlation reveals the linear relationship between features and the class labels only. This approach is insufficient and incapable of revealing the non-linear relationships.

2.2.4 Overall Comparisons

In words, random approach could promote the diversity of a MCS, however the classification accuracy of the MCS has not been considered. On the other hand, heuristic approach depends on human knowledge to the attacks and thus is easily biased. The input decimation approach uses theoretical method to find the feature groups, however the correlation coefficient ignores non-linear correlation between input features and the class labels. These deficiencies could be remedied by using the mutual information as the feature grouping criterion.

3 Framework of the MCS with Mutual Information Feature Grouping

The MCS with Mutual Information Feature Grouping (MIFG) is designed to provide a multitude of ways to increase the detection accuracy and lowering the false alarm rate. It encompasses a diverse MCS by incorporating the feature grouping approach. The superior performance of MCS provides a basis for the development of intrusion detection system. Mutual information is used for feature grouping due to its ability in revealing general dependencies between features and classes.

As for the architecture of the MCS, each base classifier is designed to detect one attack type. This design aided the addition of a new base classifier when novel attack is encountered. Furthermore, the fusion of the results from the base classifiers helps detecting novel attack. Single base classifier may provide insufficient high threshold to alarm the novel attack alone, however it may be able to alarm those novel attacks by combining several decisions made by different base classifiers.

3.1 MCS for Intrusion Detection

The MCS is adopted for intrusion detection encompasses *R* base classifiers, each corresponding to one attack class in the problem with *R* types of attacks. Each base classifier is then trained on the selected group of features which are highly correlated with the designated class. The MCS with MIFG could be applied to combine any classifiers, such as decision tree, neural networks and SVM. The Radial Basis Function Neural Network (RBFNN) will be used as the base classifier embedded in the MCS to demonstrate the MCS with MIFG technique in this paper because of its fast training algorithm enabling the learning from very large scale dataset and the intrusion detection problems usually consists of over millions of samples.

The fusion methods widely adopted in a MCS are majority voting and weighted sum [11,12]. Majority voting ignores the existence of diversity that motivates the ensemble of classifiers [13] and, in multi-class problems, two or more classes may have the same votes and thus no classification decision could be made. In contrast, the weighted sum does not have this problem and the real valued output from individual classifier provides confidence of the classification results. In [13], the authors suggested using a neural network to learn the weight for classifier fusion.

3.2 Mutual Information for Feature Grouping

Selecting groups of features for base classifiers of a MCS is based on the mutual information between the input features and the class labels. Mutual information is presented briefly in here and details can be referred to [14]. The initial uncertainty in the attack class is measured by the entropy $H(c)$. Let $P(c)$ be the probabilities for the different class labels and $c = 1,2,\ldots,K$ and K to be the number of class labels:

$$H(C) = -\sum_{c=1}^{K} P(c)\log P(c) \tag{1}$$

Given the feature vector $F = (f_1, \cdots, f_N)$, where *N* is the number of features, the average uncertainty is the conditional entropy:

$$H(C|F) = -\sum_{i=1}^{N}\left(P(f_i)\sum_{c=1}^{K}\left(P(c \mid f_i)\log P(c \mid f)\right)\right) \tag{2}$$

where *P*(*c*| *F*) is the conditional probability for class *c* given the input vector *F*. In general, the conditional entropy will be less than or equal to the initial entropy. The amount by which the uncertainty is decreased is, by definition, the mutual information *I*(*C*;*F*) between the class *c* and *F*:

$$I(C;F) = H(C) - H(C|F) \tag{3}$$

This function is symmetric with respect to *C* and *F* and can be reduced to the following expression:

$$I(C;F) = I(F;C) = \sum_{c=1}^{K}\sum_{i=1}^{N} P(c, f_i)\log\frac{P(c, f_i)}{P(c)P(f_i)} \tag{4}$$

The features being selected are those maximizing the *I(C;F)* and this indicates that a strong discriminating information is being carried by those features for a particular class. Thus those features with large *I(C;F)* are selected for a given class. Iteratively groups of features are selected for each class.

4 Experimental Results

In this section, an experiment is conducted on KDDCUP 1999 dataset which is prepared from the 1998 DARPA intrusion detection evaluation program to evaluate the performance of the proposed approach and compared with other feature grouping methods, i.e. random approach, heuristic approach and input decimation, which are mentioned in Section 2.2. In the dataset, there are substantial amount of network connections and each connection is represented by 41 features. Among the five classes of network connections, normal traffic and Denial of Service (DoS) attacks are selected for the experiment. For the DoS attack in the dataset, only 6 kinds of them will be used for experiment which are Smurf, Neptune, Back, Land, Teardrop and Ping of Death. Thus the MCS for DoS detection is built in the experiments and the methodology presented in this paper is applicable to all other types of attacks.

In the experiment of the MCS with MIFG, the features are selected according to the mutual information between input features and attack classes. The random approach selects features randomly for each of the base classifiers and 5-fold cross--validation is used. The result of the random approach is the average of the classification accuracies and false alarm rates over the 5-folds. For the heuristic approach, features are divided into 3 groups: content features, intrinsic features and traffic features based on their semantic meanings. In [3], authors suggested that features with very different meanings and related to different network traffic characteristics should be divided into two groups. On the other hand, the experimental setup for the input decimation is by selecting the features with the highest correlation coefficient to the corresponding class of the base classifier [6]. The number of base classifiers in a MCS for the proposed approach, random and input decimation is chosen as the number of attack classes in the dataset, 7 in this case and each base classifier is designated to detect one attack type, while for the heuristic approach is 3 which follows the division of the whole feature set into 3 types, each base classifier detect all kinds of attacks by examining one set of features. In each experiment, the dataset is randomly split into two half: training dataset and testing dataset. The testing dataset serves as future unseen samples to evaluate the final result of the different methods and the training of the MCSs use only the training dataset in all the methods. All the input values are normalized to [0,1] and weighted sum is used as the fusion method for combining the results of base classifiers using different feature grouping approaches.

The performances of the MCSs with different feature grouping methods for detecting DoS attacks are shown in Table 1. The MCS built using the input decimation attains unsatisfactory testing accuracy and relatively high false alarm rate. This indicates that input decimation is not suitable for detecting DoS attacks in which non-linear correlation exists between features input and output classes. This non-linear correlation could carry valuable information and the ignorance of these information results in information loss. Thus input decimation is only suitable for the problem which possesses linear relationship between input and output. On the other hand, the performances of the MCSs built using the random and heuristic approaches are comparable, however both of them have 4% less in testing accuracy and 4% more in false alarm rate when compared with the proposed method. The heuristic approach will be perfect only if the relationship between inputs and outputs are fully understood by the domain expert but this is usually not true. In [3], the performance using heuristic approach is satisfactory. However it only considers the ftp service which may be due to the domain knowledge of the expert is being stronger in such cases, but not for all the attacks. Different problems require different experts to define the feature set which is problematic and they may ignore some hidden relation between input features and the attacks while overestimate the importance of some input features. The MCS built using the random approach performs better than the one using the input decimation and heuristic approaches. However the performance of the random approach is unstable and the classification accuracy has not been considered.

The MCS built using the proposed MIFG outperforms the MCSs built using other feature grouping methods in the experiment. There is 2% to 7% increase in testing accuracy when adopting the proposed approach to compare with other approaches. For the false alarm rate, the proposed approach shows only 1% while there is 3% to 8% for the other methods. Thus the proposed approach provides satisfactory performance with high accuracy and low false alarm rate which is the main goal of an effective IDS [15,16]. Mutual information is shown to possess the ability in revealing the non-linear relationship between input features and class labels theoretically and provides stable and non-biased performance in detecting DoS attacks which is further supported by the experimental results.

Table 1. Experimental Results for MCS using different feature grouping methods on the KDDCUP'99 Dataset

Feature Grouping Methods	Testing Accuracy	False Alarm Rate
MIFG	***98.20%***	***1.28%***
Random	94.29%	5.31%
Heuristic	94.17%	5.82%
Input decimation	91.00%	8.99%

5 Conclusion and Future Works

In this paper, we proposed to use the mutual information in feature grouping (MIFG) for MCS training. The MIFG promotes diversity of base classifiers in the MCS and reveals the hidden relationship between the input features and the class labels (classifier output). A brief survey of the current feature grouping methods is given. Network intrusion detection problem is discussed and we applied the MCS with MIFG in the experiments to detect the intrusion. Experimental results are promising and the MCS built using MIFG outperforms other current feature grouping methods.

The characteristics of the problems that benefit from the MCS with MIFG approach have been examined. The proposed approach performs best for large sample size and the intrusion detection problem which tackles a large number of connection records is an example. A complex problem with many classes can also benefit from the proposed approach since it will be divided into smaller and simpler problems. Simplifying the learning tasks allow classifiers to achieve better performance.

Acknowledgment

This work is supported by the Hong Kong Polytechnic University Research Grant GT-891.

Reference

1. L. Xu, A. Krzyzak and C. Y. Suen, “Methods for combining multiple classifiers and their applications to handwriting recognition”,IEEETrans.SMC,vol.22,pp. 418-435, 1992.
2. J. Kittler, M. Hatef, R. P. W. Duin and J. Matas, “On Combining Classifiers”, IEEE Trans. On Pattern Analysis and Machine Intelligence, 20, pp. 226-239, 1998.
3. G. Giorgio, R. Fabio & D. Luca, “Fusion of multiple classifiers for intrusion detection in computer networks”, Pattern Recognition Letters 24, pp. 1795-1803, 2003.
4. S. Mukkamala, A. H. Sung and A. Abraham, “Intrusion detection using an ensemble of intelligent paradigms”, Journal of Network and Computer applications, 2004.
5. L. Hanson and P. Salamon, “Neural Network Ensembles”, IEEE Transactions on pattern analysis and machine intelligence, 1990.
6. N. C. Oza and K. Tumer, “Input Decimation Ensembles: Decorrelation through Dimensinality Reduction”, MCS 2001, LNCS 2096, pp. 238-247, 2001.
7. K. Tumer, J. Ghosh, Classifier combining: analytical results and implications, Working notes from the Workshop ‘Integrating Multiple Learned Models’, 13th National Conference on Artificial Intelligence, August 1996, Protland, Oregon.
8. T. K. Ho, “The random subspace method for constructing decision forests”, IEEE Transactions on pattern analysis and machine intelligence, 1998.
9. R. Bryll, R. Gutierrez-Osuna, F. Quek, “Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets”, Pattern recognition 2003.
10. L. I. Kuncheva, Genetic algorithm for feature selection for parallel classifiers, Information Processing Letters, pp. 46:163-168, 1993.
11. T. H. Ho, J. J. Hull and S.N. Srihari, “Decision Combination in Multiple Classifier System”, IEEE Trans. PAMI, vol. 16, pt. 1, pp. 66-75, 1994.
12. R. R. Yager and J. Kacprzyk, editors.,“The Ordered Weighted Averaging Operators”, Kluwer Academic Publishers, 1997.
13. W. W. Y. Ng, A. P. F. Chan, D. S. Yeung and E. C. C. Tsang, “Quantitative Study on the Generalization Error of Multiple Classifier Systems”, Submitted to IC-SMC 2005.

14. R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning", IEEE Transactions on neural networks, Vol 5, No. 4, July 1994.
15. S. Axelsson, "The base-rate fallacy and the difficulty of intrusion detection", ACM Trans. Information Syst. Security 3 (3), pp. 186-205, 2000.
16. P. E. Proctor, "The Practical Intrusion Detection Handbook", Prentice Hall, 2001.
17. S. Kumar, EH. Spafford, "A pattern matching model for misuse intrusion detection", Proceedings of the 17th National Computer Security Conference, 1994b.
18. W. Lee, S. J. Stolfo, "Data mining approaches for intrusion detection", Proceedings of the seventh USENIX Security Symposium, 1998.
19. R. Schapire, "The strength of weak learnability ", Machine Learning, pp. 197-227, 1990.
20. T. Verwoerd, R. Hunt, "Intrusion detection techniques and approaches", Computer communications, pp. 1356-1365, 2002.
21. W. W. Y. Ng, R. K. C. Chang and D. S. Yeung, "Dimensionality Reduction for Denial of Service Detection Problems Using RBFNN Output Sensitivity", Proc. on International Conference of Machine Learning and Cybernetics, pp. 1298 – 1298, 2003.

Design and Implement a Web News Retrieval System

James N.K. Liu, Weidong Luo, and Edmond M.C. Chan

Department of Computing,The Hong Kong Polytechnic University
{csnkliu,cswdluo}@comp.polyu.edu.hk, spmcchan@inet.polyu.edu.hk

Abstract. We present the design and implementation of "Ai-Times", a web-based news retrieval system. We also describe the spider module, news extraction module and automatic summarization module in detail.

1 Introduction

The explosion in the availability of online information easily accessible through the Internet is a reality. As the available information increases, the inability to process, assimilate and use such large amounts of information becomes more and more apparent. Online news information suffers from these problems. In addition, currently available search engines present inefficient behavior, for example, cannot offer users the facility of specifying the categories and time frames they receive and cannot provide the online news information with required frequency.

We describe the design and implementation of "Ai-Times", a web-based news retrieval system the goal of which is to accurately retrieve and organize the web news information. We also describe the spider module, news extraction module and automatic summarization module in detail. We believe many existing news retrieval systems have already used some of these techniques, but there has been little scientific research conducted on these techniques.

The remainder of this paper will be organized as follow: In section 2 we survey the related work on news retrieval systems. In section 3 we define key terms and describe the operation of Ai-Times. We also describe the optimized web spider algorithm, the news content extract module and the automatic summarization module in detail. In section 4 we give experimental result. The final section offers our conclusion.

2 Related Work

News information retrieval has been well studied [1, 2, 3, 4, 5]. Many studies have been done on the general architecture of news information retrieval. For example, reference [1] presents a system that automatically classifies TV news articles using keywords; Reference [2] describes the THISL news retrieval system maintaining an archive of BBC radio and televisions news records; Reference [3] introduces the WIRE-a WWW-based information retrieval and extraction system; Reference [4] presents the design of a news retrieval tools based on an existing database of some newspapers such as Times. The above literatures contribute much to the general architecture of news information retrieval. However, none of them describes the core modules such as the spider module, the news extraction module and the automatically summarization module in detail.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 149–156, 2005.

There also exists some literature studying the data-rich part extraction algorithm and web spider algorithm. For example, Reference [5] describes an algorithm for automatically partitioning the html document into tree-like semantic structures, which expose the implicit schema. However, it can't perform reasonably well on all types of html pages when used in news extraction. Reference [8] introduces how to implement an effective web spider. However, the general-purpose spider's performance is relatively poor when it is directly applied to news retrieval problem.

3 The Architecture and Algorithm of Ai-Times

In this section we describe the basic framework of the Ai-Times. We also describe the spider module, news extraction module and summarization module's algorithm in detail.

3.1 Architecture

Figure 1 shows the architecture of the Ai-times system, which contains several classic web based information retrieval system modules such as a web spider, an auto-categorization module, an index engine, a search module, and an auto-summarization module. The specialty of the Ai-times lies in these: Ai-times can extract the news title, text content and pictures automatically with very little manual job, Ai-Times optimizes the web spider algorithm to save the crawler and update time, Ai-times can also provide the summarization of news.

We discuss the web spider module and the news-extraction module and the automatic summarization module in detail from section 3.2 to section 3.4.

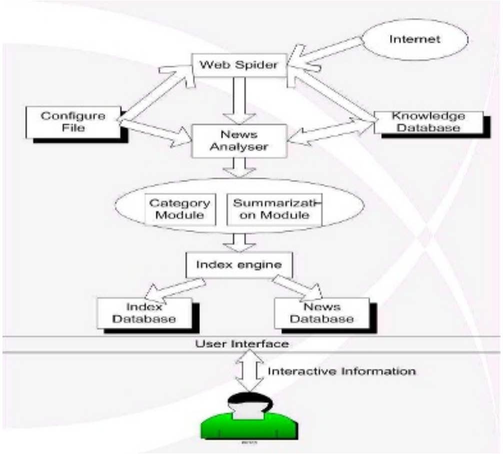


Fig. 1. The architecture of the Ai-Times news retrieval system

3.2 Web Spider

One of the fundamental and important components of news information retrieval system is the Web Spider that can collect the web document automatically. Much research has been done in this area, for instance, the Cobweb [6], is a typical Web Spider. Ai-Times focuses on the news collecting from some predefined news web sites, so the Ai-Times web spider algorithm is different from the classic spider algorithm. The classic spider algorithm has been discussed in literature [7], we present an optimized algorithm that is applicable to the web news spider.

The definitions of valueless web document, news content web document, index or list web document are given below:

Valueless Web Document: The web documents that contribute nothing to the news retrieval, for example, the advertisement page.

News Content Web Document: web news documents that mainly contain the news text content or the news images or other multimedia sources.

Index or List Web Document: web documents that mainly contain hyperlinks linking to the News content web document with relevant caption; often, the caption is the news title.

Algorithm I: Optimized Spider Algorithm

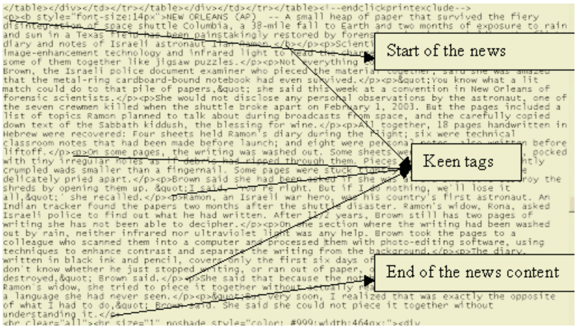
```
Begin
Let I be a list of initial URLs of the news website;
Let F be a queue;
   For each URL i in I
       Enqueue(i,F);
   End
   While F is not empty
       u=Dequeue(F);
       if u has not been processed
           Get (u);
           Case u's type:
           Valueless web document:
                   Skip u.
           News content web document:
                   Store u;
           Index or list page:
                   Extract the hyperlinks and relevant caption;
                   Let U be the set of hyperlinks extracted;
                   For each hyperlink u in U
                        Enqueue(u,F);
                   End
       Else—u has already been processed
           Case u's type
           Valueless web document:
                   Skip u;
           News content web document:
                   Skip u;
           Index or list web document:
                   Update checking
      End
   End
End
```

It is often a time-consuming job to check whether a web document has been modified or updated. In general, the spider should often rescan all the websites and all web documents for update checking. Sometimes, the spider sends request to the web server and analyzes the return http header from the web server to get the last modified time of the web document. Then, the spider will determine whether to refresh the downloaded web document or not. However, even the spider can get the last modified time through the HTTP header, it still waste much request time and system resources. Nevertheless, many HTTP servers do not provide the last modified time in the HTTP header. Therefore, one of the common refresh policies is to revisit all web documents to find the update information, however, this is very time-consuming. Another refresh policy is to select some important pages to revisit and this method often loses information [7].

Generally, more than 90% web documents of a news web site are news content web documents and these web documents are seldom or never modified or updated. However, the index or the list of the web documents is updated frequently. By defining the three kinds of web documents, the Ai-Times web spider need not request all web documents during the update interval. As we can see from the Algorithm I, our spider will ignore the valueless web document and news content web document and only revisit the index or list web document for update checking. This saves much time and makes the refresh interval shorter.

3.3 News Extraction

In this section we introduce a new concept what we called “keen tags”. We also describe how to extract the news text content using the “keen tags”. By investigating a large number of news web sites, we found that in most news web documents, the news text content string tokens divided by the html tokens will spatially cluster together in the html source code. In order to extract the news text content using a single general-purpose extractor instead of using different website-related extractors, here we will introduce a concept of “Keen Tag”. A keen tag is a tag that always appears inside or around the news text content. Figure 2 shows a “keen tags” example from http://www.cnn.com.



```
</table></div></td></tr></table></div></td></tr></table><!--endclickprintexclude-->
<p><b style="font-size:14px">NEW ORLEANS (AP)  -- A small heap of paper that survived the fiery
disintegration of space shuttle Columbia, a 38-mile fall to Earth and two months of exposure to rain
and sun in a Texas field has been painstakingly restored by forens
diary and notes of Israeli astronaut Ilan Ramon.</b></p><p>Scienti
image-enhancement technology and infrared light to read the charr
some of them together like jigsaw puzzles.</p><p>Not everything co
Brown, the Israeli police document examiner who pieced the material together, said she was amazed
that the metal-ring cardboard-bound notebook had even survived.</p><p>&quot;You know what a lit
match could do to that pile of papers,&quot; she said this week at a convention in New Orleans of
forensic scientists.</p><p>She would not disclose any personal observations by the astronaut, one of
the seven crewmen killed when the shuttle broke apart on February 1, 2003. But the pages included a
list of topics Ramon planned to talk about during broadcasts from space, and the carefully copied
down text of the Sabbath kiddush, the blessing for wine.</p><p>All together, 18 pages handwritten in
Hebrew were recovered: Four sheets held Ramon's diary during the flight; six were technical
classroom notes that had been made before launch; and eight were personal notes, also written before
liftoff.</p><p>On some pages, the writing was washed out. Some sheets we                  , pocked
with tiny irregular holes as if debris had ripped through them. Pieces                   ntly
crumpled wads smaller than a fingernail. Some pages were stuck tight                     e
delicately pried apart.</p><p>Brown said she had been asked if she was a                  roy the
shreds by opening them up. &quot;I said, 'You're right. But if I do nothing, we'll lose it
all,&quot;' she recalled.</p><p>Ramon, an Israeli war hero, was his country's first astronaut. An
Indian tracker found the papers two months after the shuttle disaster. Ramon's widow, Rona, asked
Israeli police to find out what he had written. After 11/2 years, Brown still has two pages of
writing she has not been able to decipher.</p><p>On one section where the writing had been washed
out by rain, neither infrared nor ultraviolet light was any help. Brown took the pages to a
colleague who scanned them into a computer and processed them with photo-editing software, using
techniques to enhance contrast and separate the writing from the background.</p><p>The diary,
written in black ink and pencil, covers only the first six days of
don't know whether he just stopped writing, or ran out of paper, o
destroyed,&quot; Brown said.</p><p>She said that because the not
Ramon's widow, she tried to piece it together without actually rea
a language she had never seen.</p><p>&quot;But very soon, I realized that was exactly the opposite
of what I had to do,&quot; Brown said. She said she could not piece it together without
understanding it.</p>
<br clear="all"><hr size="1" noshade style="color: #999;width:464px;"><div
```

Fig. 2. A source example of the news content of a typical web document

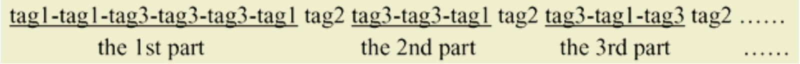
As we can see from Figure 2, the news content strings spatially cluster together in the html source and are accompanied with the Tag “<p>” and “</p>”. Universally, the tags such as “<p>”,” </p>”,”
”,” <font>”,” <img..>” can always be a web site’s keen tag. We define a list containing universal keen tags: <p>, </p>, <font*>, </font>, <b>, <a href*>, </a>, <img*>, <br>, <strong>, </strong>, </div>, <div>, <center>, </center>.

By defining the keen tag, we divide the string tokens of the html source of the web document into 3 types:

1. Keen tags. 2. Html tags that is not a keen tag. 3. String tokens that are not html tag.

Algorithm II: News Extraction Algorithm

1. Separate the html source code into many parts by the type 2 tags; therefore, each part contains the type1 and type3 strings. After that, the html source of the news document would look like this:


tag1-tag1-tag3-tag3-tag3-tag1 tag2 tag3-tag3-tag1 tag2 tag3-tag1-tag3 tag2 ......
the 1st part
the 2nd part
the 3rd part
......

2. Score each part, the scoring formulation is:

$$\text{Score (i)}=\sum_{k=1}^{ni} len(k) \qquad (1)$$

where i is the sequence of the part. For i=1,2,…n (n is the total number of the part), ni is the total number of type3 tags in the current part, k is the sequence of the type3 tags in the current part, len(k) is the length of the kth type3 tag in the current part for k=1, 2, …. ni.

3. Select the score winner part as the data-rich part.
4. Evaluate the selected part to see if it is a news content web document:

$$\text{evaluate()=fin(score)* tin(hrenum)} \qquad (2)$$

where score is the score obtained from equation 1, hrefnum is the total number of hyperlink html tags(html tag like<a href…>) in that part.

$$\text{fin(k)=1 while k>N, and fin(k)=0 while k} \le \text{N}$$

where N is the predefined length of the news content. Any news document whose content length is shorter than N will not be considered to be a useful news web document.

$$\text{tin(k)=1 while k<T tin(k)=0 while k} \ge \text{T}$$

where T is the predefined up range of the total number of the link tags in a news content. Any news document that contains more than T link tags within the news content will not be considered a useful news web document.

5. If evaluate() = 1, the document can be regarded as a news content web document, and then the selected data rich part will be extracted as the news text content. Let the caption we got before to be the news' title. Let current time to be the download time of this news.

Besides the universal keen tags we have defined before, there are still some website-relevant keen tags that we can derive from the training phrase. At the training phrase, the well-defined web news document will be inputted and the system will analyze these web news documents based on some pre-defined rules, then output the website-relevant keen tags.

3.4 Automatic News Summarization

In this section we describe the automatic summarization module of Ai-Times. The process of Ai-Times automatic summarization consists of following three steps: characteristic words generation, sentence weighting and summary generation. In section 3.4.1 we describe the characteristic words generation. In section 3.4.2 we describe the sentence weighting and in section 3.4.3 we describe the summary generation.

3.4.1 Characteristic Words Generation

Characteristic words are generated by word segmentation algorithm and characteristic word weighting equation based on Statistical and Probabilistic Approach. Words in web news document are assigned to different weighting based on their importance and frequency in document. Words with highest weighting are extracted as characteristic words.

The summarization algorithm scans a web news document and finds out proper words in the web news document using the following equation:

$$P(w) = F_t(w) \bullet \left(1 - \frac{numdoc}{tota \ln umdoc}\right) \bullet \left(L(w) - D\right)^c$$

where w is the Chinese word extracted from a sentence in the web news document

$F_t(w)$ is the frequency of w, L(w) is the length of w
numdoc is the number of documents which contains w
totalnumdoc is the total number of documents
D is the minimum length of w
c is the influence of the length of the sequence of Chinese characters

The characteristic words can only be treated as keywords for retrieving the relevant web news documents. These words carry no semantic meaning.

3.4.2 Sentence Weighting

So far, we have described how to generate characteristic words. Note that the second step of our summarization algorithm is sentence weighting. In this step sentences are assigned weights, which indicate their importance in a web news document. The higher the weight, the greater the importance. The highest weight is assigned to the sentence that presents the most important concept of the web news document. Important sentences must have the features of more characteristic words in the sentence; of higher characteristic weight of characteristic words in the sentence; of shorter length of sentences; of having less sub-sentences for an important sentence; and of less numeral words in the sentence. The following equation computes the sentence's importance weight [8].

$$T(s) = \frac{\sum_{i=t}^{N} Ti}{s_0 \quad \bullet \quad s_1 \quad \bullet \quad s_2 \quad \bullet \quad m}$$

where N is the number of Characteristic words in the sentence.

Ti, i=1 to N, is characteristic weight of the ith characteristic word in the sentence.
s0 is the total number of words in the sentence.
s1 is the number of sub-sentences in the sentence.
s2 is the number of numeral words in the sentence.
m is an integer variable. Normally, it is set to 1.

The sentence containing more characteristic words will be assigned a higher weight. The sentence containing more number of sub-sentences, words and numeral will be assigned a smaller weight. Important sentences can then be extracted from a document.

3.4.3 Summary Generation

The third step of Ai-Times summarization algorithm is summary generation. In this step, the sentences with the highest sentence weight are extracted to generate a summary in a particular ratio required by the user according to the following equation [8].

$$Abstract = \sum_{i=1}^{x} |s_i| = T \bullet R$$

where T is the length of the web news document.

R is the abstract ratio.
si is the sentence with ith highest sentence weight.
| si | is the length of si.

4 Experiment

Effectiveness is the main focus for the overall system. On average, Ai-times processes 10 web news documents per second (processing include download, store, index and generate summary) on a single common pc server. The server hardware configuration is: CPU: p4 2.4 GB; RAM: 512 KB; Bandwidth: 1MB

In total it took roughly 6 days to process 3.53 million web pages (including errors and the valueless web documents). By using the Web Spider and News Extract (and abstraction) Module, 1.46 million news documents were downloaded from 50 online news websites and news content including the pictures were extracted and stored for indexing and retrieval.

91.3% news content and pictures extracted are correct and 13 percents of the valuable news documents are missed. The performance is acceptable because we have only implemented a news spider instead of having different spiders and extractors for each web site.

Note that we have defined three types of web news document (defined in section 3.1), Our spider needs not revisit all the web documents during refresh interval. This policy saves much time and network traffic. In the experiment, our spider obtained 8 times improvement compared to traditional spider during the refresh period.

5 Conclusion

We have presented a web based news retrieval system and introduced an optimized web spider algorithm. We also provided a “Keen tag” analyzing method for extracting the news content from the news web document. Our other contribution is that we have presented an automatic summarization algorithm that can generate news summary automatically.

In summary, we believe that our paper offers some useful guideline for the news information retrieval system designer, helping them for example, optimize their spider to be more time-saving, implement a more flexible news content extractor, or implement an applicable news summarization module.

Acknowledgement

The authors would like to acknowledge the partial support of the ICRG grant A-PF83 and Teaching Company Scheme between Hong Kong Polytechnic University and Ai-Times Company Ltd.

References

1. Yasuo Ariki, Yoshiaki Sugiyama.A TV News Retrieval System with Interactive Query Function. Proceedings of the Second IFCIS International Conference on Cooperative Information Systems(1997)184-192.
2. Steve Renals, Dave Abberley, David Kirby and Tony Robinson .The THISL system for indexing and retrieval of broadcast news. IEEE Signal Processing Society 1999 Workshop on Multimedia Signal Processing September 13-15, 1999, Copenhagen, Denmark(1999) 77-82.
3. Sudhir Aggarwal, Fuyung Hung. WIRE - A WWW-based Information Retrieval and Extraction System. Proc. of the 9th International Workshop on Database and Expert Systems Applications (DEXA'98), Vienna, Austria.(1998)887-892.
4. M. Sanderson & C.J. van Rijsbergen. NRT - News Retrieval Tool. Electronic Publishing, EP-odd, Vol. 4, Num. 4(1991) 205-217
5. Saikat Mukherjee, Guizhen Yang, Wenfang Tan, I. V. Ramakrishnan. Automatic Discovery of Semantic Structures in HTML Documents. ICDAR (2003) 245-249.
6. CoBWeb – A Crawler for the BrazilianWeb, Proceedings of the String Processing and Information Retrieval Symposium & International Workshop on Groupware table of contents (1999) 184.
7. Cho, Junghoo, Garcia-Molina, Hector. Effective Page Refresh Policies for Web Crawlers, ACM Trans. Database System. 28(4) (2003) 390-426.
8. Li, J.J.“Research and Implementation of A Domain-Unconstrained Chinese Automatic Abstracting System”Phd dissertation, Dept. of Computer Science, Harbin Institute of Technology. 1996.

An Ontology for Integrating Multimedia Databases

Chull Hwan Song[1], Young Hyun Koo[1], Seong Joon Yoo[1,*], and ByeongHo Choi[2]

[1] School of Computer Engineering, Sejong University,
98 Gunja, Gwangjin, Seoul, Korea 143-747
sjyoo@sejong.ac.kr
[2] Digital Media Research Center, KETI, 270-2,
Seohyun-Dong, Pundang-Gu, Sungnam-Si, KyungGi-Do, Korea

Abstract. This paper proposes a multimedia ontology that will support semantic integration of multimedia databases. The ontology is defined by integrating MPEG-7 MDS and TV-AnyTime metadata. A procedure to integrate these two standards is described and relationship of their elements is also presented. We also designed and implemented a framework for integrating multimedia databases and showed that any multimedia metadata can be mapped to the proposed ontology.

1 Introduction

This paper aims at defining a multimedia ontology for semantic integration of multimedia databases (hereafter we call it SM ontology), which can search various types of multimedia data in distributed heterogeneous environments. Defining SM ontology starts from ensuring compatibility with the current international multimedia metadata standards. In other words, it is necessary to ensure compatibility mainly with MPEG-7 Multimedia Description Scheme (MDS) [1] and TV Anytime metadata [2]. This SM ontology can be used for the basis of designing federated multimedia information retrieval system on the semantic web.

For this purpose, MPEG-7 MDS and TV-Anytime metadata were compared and analyzed to identify commonness and differences between these two standards. Based on the results, a multimedia ontology made up of a union of these two standards was created. Investigating these two standards revealed that some definitions were expressed in different terms even though they have the same meaning. In this case, either ones were used or third terms were selected. In another case, an element or an attribute of a definition in one standard included the entire elements or attributes of a definition of the other standard, which was easily integrated. When two elements had an intersection with some differences, the intersection was defined with a newly adopted term while the remnants were defined with the terms originally used in the standard.

2 Related Works

Most recent research [4][5][6][7][8] proposes ontology based approach for multimedia information systems. Especially, Tsinaraki et al.[7][8] proposes a framework

[*] Author for Correspondence

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 157–162, 2005.

which allows transforming OWL ontology to MPEG-7 and TV-Anytime. However, they have not showed an integrated metadata with which access MPEG-7 and TV-Anytime metadata. Tsinaraki et al. transforms OWL to MPEG-7 and OWL to TV-Anytime respectively. The approach of transforming OWL to each metadata is not flexible enough to support non-standard metadata. This paper proposes a universal metadata that will support non-standard multimedia metadata as well as the standard multimedia metadata.

3 Procedures to Integrate MPEG-7 MDS and TV-AnyTime Metadata

Of these two standards, MPEG-7 MDS deals with all digital contents such as image information, audio, video, and 3D as well as motion images, unlike TV-AnyTime Fig. 1 shows the relationship between MPEG-7 MDS and TV-AnyTime. Most elements of TV-AnyTime metadata correspond to some elements of MPEG-7 MDS.

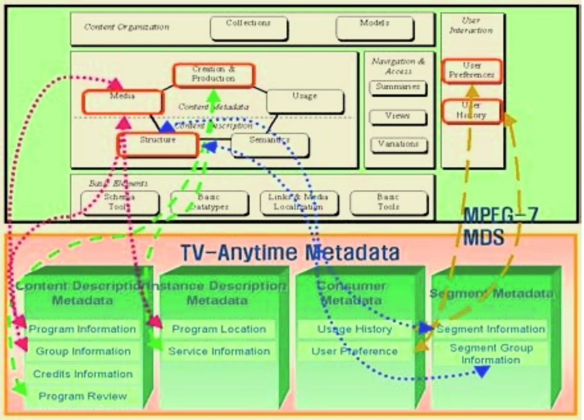


Fig. 1. Relationship of MPEG-7 MDS and TV-AnyTime Metadata

The two standards are integrated through four steps. The first step is to investigate the similarities of their high level structures, which contributes to establishing the relation of inclusion between them. The second step is to examine the relationship among low-level elements of the higher structures. The low-level elements of one standard could be high-level elements of the other standard. The results of the first and the second steps revealed a certain relation of intersection between the two standards, according to which their low-level elements were divided into four groups to clarify their relationship. In this process, relationship of inclusion and intersection between the two standards were established. The standard that occupied a large part in the relationship of inclusion became the standard for the proposed metadata integration method. For the relationship of intersection, if basic concepts of high-level elements were similar and those of their low-level elements were also similar, all of them were included in the standard for the proposed method. On the other hand, if concepts of high-level elements were similar but their low-level elements were divided into two or three different concepts, two or three different high-level elements were established in the standard for the proposed method, to include each of low-level elements. The third step is to construct a mapping table based on the results of the investigation of the relationship of the two standards. The mapped elements had element names common to both standards. The fourth step is to create new ontology that integrates all elements as well as those mapped elements.

4 SM Ontology

While most of the elements of the MPEG-7 MDS and TVA metadata can be mapped directly to SM ontology, common part such as MediaFormatType and AVAttributes, VisualCoding and VideoAttributes, AudioCoding and AudioAttributes, CreateInformationType and BasicDescription, and UsageHistory and UserPreferenceType of these two standards need to be described in this section. We, however, could illustrate only a part of the proposed multimedia ontology in this section due to space limit.

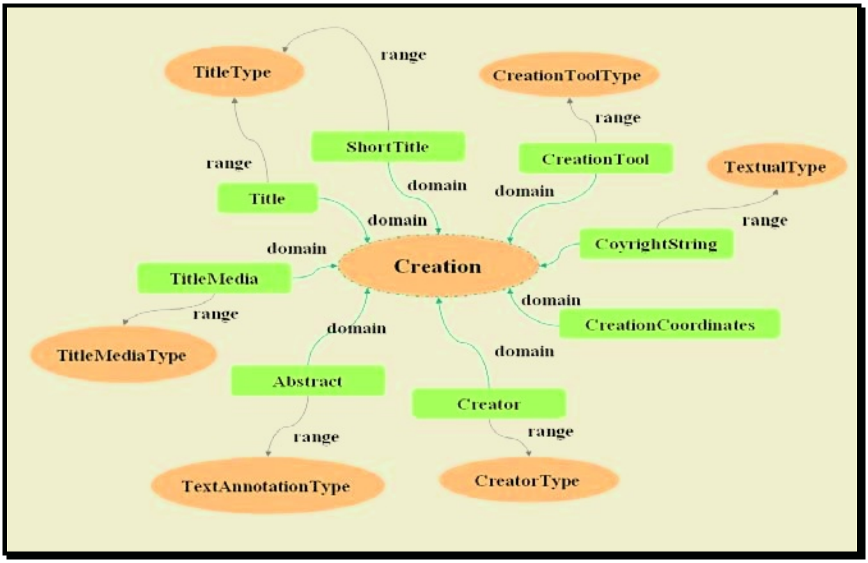


Fig. 2. SM Ontology for Creation Information

Fig. 2 depicts a common ontology for multimedia creation information. The creation concept is described by title, abstract, creator, creation tool, copyright information and creation coordinates. This ontology for creation information is defined by integrating CreationType of MPEG-7 MDS and a part of BasicDescription of TV-AnyTime metadata.

Fig. 3 illustrates an ontology for creation information of multimedia data. Genre, Language, CaptionLanguage, SignLanguage, Release and ParentalGuidance of BasicDescription element of TV-AnyTime metadata correspond to some elements of Classification element of MPEG-7 MDS. Classification is composed of Genre, Form, Purpose, Keyword, Target, ParentalGuidance, Release, Media Review and Language information.

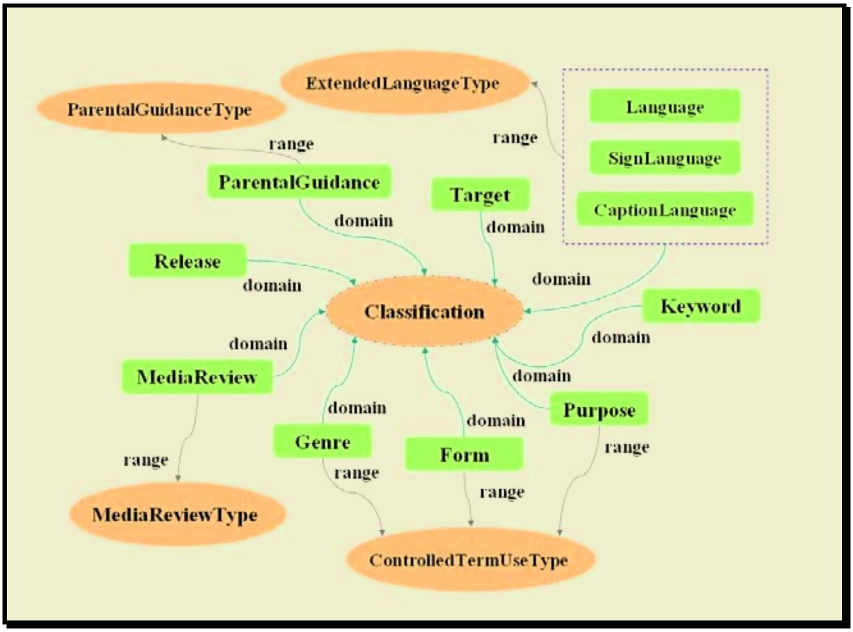


Fig. 3. SM Ontology for Multimedia Classification

5 Implementation and Experiment of a Web Service Based Multimedia Integration System

This section describes the architecture of the web service based multimedia integration system that we have implemented for experimenting the applicability of UMA metadata. Since the integration system is implemented using web services technology, it gives users more flexibility than previous multimedia integration systems. The integration system is composed of three layers as shown in Fig. 4: application layer, mediation layer, and resource layer. The web service API provided by the mediation layer and the resource layer are used for data transfer between layers.

5.1 Application Layer

In Application Layer, a user or an upper module transfers query and receives the results. For example, a user can generate in simple and general query such as name, genre, ID, or keyword. Application Layer calls the web service API provided by Mediation Layer. This API transfers the query to Mediation Layer in XML format using SOAP protocol. Also, this API transfers the results of query to Application Layer in XML format.

Fig. 5 shows an application system with which users can access video data distributed over network and tagged with various metadata. This application system is implemented using web service API provided by the prototype system.

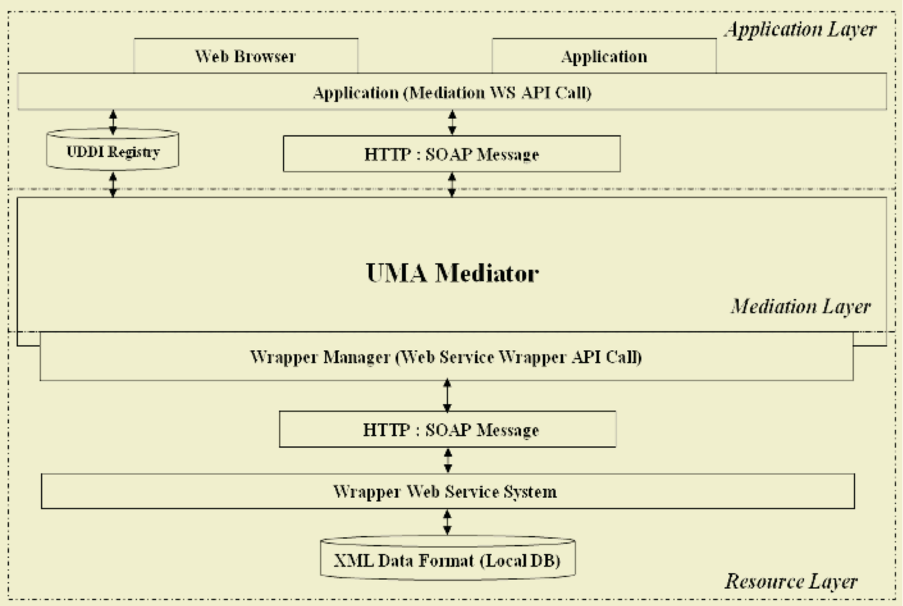
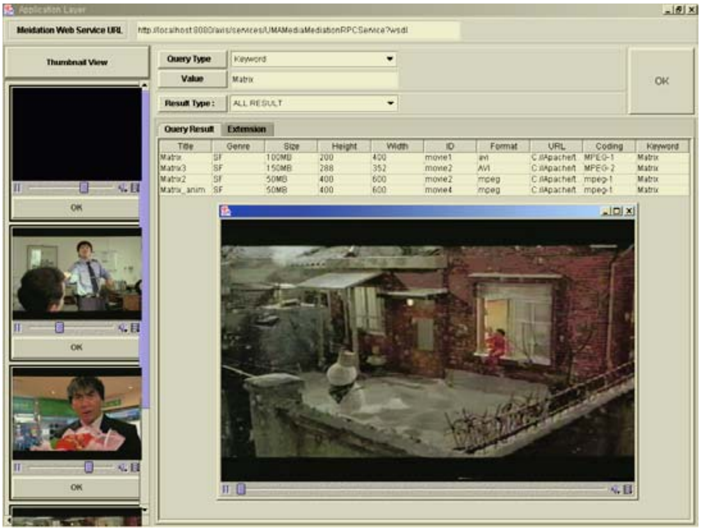

Application Layer
Web Browser
Application
Application (Mediation WS API Call)
UDDI Registry
HTTP : SOAP Message
UMA Mediator
Mediation Layer
Wrapper Manager (Web Service Wrapper API Call)
HTTP : SOAP Message
Wrapper Web Service System
XML Data Format (Local DB)
Resource Layer

Fig. 4. Web Service Based Integration of Multimedia Databases



Meidation Web Service URL: http://localhost:8080/axis/services/UMAMediaMediationRPCService?wsdl

Query Type: Keyword

Value: Matrix

Result Type : ALL RESULT

Title	Genre	Size	Height	Width	ID	Format	URL	Coding	Keyword
Matrix	SF	100MB	200	400	movie1	avi	C:\\Apache\	MPEG-1	Matrix
Matrix3	SF	150MB	288	352	movie2	AVI	C:\\Apache\	MPEG-2	Matrix
Matrix2	SF	50MB	400	600	movie2	mpeg	C:\\Apache\	mpeg-1	Matrix
Matrix_anim	SF	50MB	400	600	movie4	mpeg	C:\\Apache\	mpeg-1	Matrix

Fig. 5. SM Ontology based Integrated Video Data Retrieval System

5.2 Mediation Layer

Mediation Layer is composed of Query Processor, Rule Manager, and Global Schema Manager. Query is transferred from Application Layer to Mediation Layer using the API methods.

5.3 Resource Layer

Resource Layer is connected to Local DB through wrappers. This layer retrieves relevant data from Local DB. Resource Layer is composed of a wrapper manager and multiple wrappers. Since the major functions are provided with web services API, users can build their own wrapper with ease.

6 Conclusions

This paper introduced a new metadata for multimedia contents. In order to define and construct the new metadata, two standards, which are internationally recognized to have the most multimedia information, were compared to investigate their relationship. In order to prove the effectiveness, we have implemented a prototype system for integrating multimedia databases. We have tested UMA data by integrating databases tagged with five different standards and found that the mapping tool enables defining the relationship between UMA and the five standards semi-automatically. In addition, users can build their own wrappers and compose the relationship between wrappers and a mediator more easily than ever since the prototype system provides web service API. We have showed this efficiency by showing easily implemented prototype system in the previous section. In the future, this metadata for multimedia contents will be used in defining multimedia ontology language, like RDF or OWL [9][10], so that it can be used for context aware multimedia access.

References

1. Martinez, J. M.: Overview of the MPEG-7 Standard (version 5.0): ISO/IEC JTC1/SC29/WG11 N4031, Singapore, March (2001), http://www.cselt.it/mpeg/standards/mpeg-7/mpeg-7.htm
2. TV-Anytime Forum, http://www.tv-anytime.org/
3. Roantree, M.: Metadata Management in Federated Multimedia Information Systems: Proceedings of the thirteenth Australasian Conference on Database Technologies. Vol. 5 (2002) 147-155
4. Hunter, J.: Adding Multimedia to the Semantic Web-Building an Mpeg-7 Ontology: In nternational Semantic Web Working Symposium (SWWS), Stanford, (2001) 261-283
5. Chotmanee, A.; Wuwongse,V., Anutariya,C.: A Schema Language for MPEG-7: LNCS 2555, Vol. 2555 (2002) 153-164
6. Troncy, R.: Integrating Structure and Semantics into Audio-visual Documents: In 2nd International Semantic Web Conference (ISWC'03), LNCS 2870, Sanibel Island, Florida, USA, Vol. 2870 (2003) 566-581
7. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Integration of OWL ontologies in MPEG-7 and TVAnytime compliant Semantic Indexing: In the proceedings of the 3rd HDMS, Athens, Greece, Vol. 3084 (2004) 398-413
8. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Interoperability Support for Ontology-Based Video Retrieval Applications: LNCS 3115, Vol. 3115 (2004) 582-591
9. OWL Web Ontology Language Reference, http://www.w3.org/TR/2004/REC-owl-ref-20040210/#EnumeratedDatatype
10. Ferdinand, M., Zirpins, C., Trastour, D.: Lifting XML Schema to OWL: in Web Engineering - 4th International Conferences, Vol. 3140 (2004) 354-358

Integrating Service Registries with OWL-S Ontologies*

Kyong-Ha Lee[1], Kyu-Chul Lee[1], Dae-Wook Lee[2], and Suk-Ho Lee[2]

[1] Department of Computer Engineering, Chungnam National University, 220,
Gung-dong, Yuseong-Gu, Daejeon, South Korea
{bart,kclee}@cnu.ac.kr

[2] School of Electric Engineering & Computer Science, Seoul National University,
Kwanak-Gu, Seoul, South Korea
dwlee@dbmain.snu.ac.kr, shlee@snu.ac.kr

Abstract. WWW is currently undergoing a remarkable change from a collection of pages to a collection of services that interoperate through the Internet. The representative types of these services are XML Web Services and ebXML, which are emerging as significant e-Business application frameworks. Both ebXML and Web Services, however, use different registries for publishing and discovery of services, and this has given rise to some problems when adopting frameworks. Furthermore, these registries support only keyword-based search, which does not make use of semantic information, and which does not address the problem of matching service capabilities and allowing service location based on the functionalities sought. This paper shows that service descriptions in OWL-S can be mapped into UDDI and ebXML registry entries, and can be migrated from registries; therefore, it provides not only a way to integrate registry entries but also a way to build the foundation of semantic service discovery. Also, migrated OWL-S can be updated periodically while registry entries are being modified.

1 Background and Motivation

While WWW is undergoing a remarkable change from a collection of pages to a collection of services that interoperate through the Internet, much effort has been exerted to utilize WWW in order to make global e-Business a reality. The dormitory standards in the area of e-Business today are Web Services and ebXML. Web Services is a *de facto* standard designed to support interoperable software components interaction over the Internet, based on XML technologies. On the other hand, ebXML is a non-profit standard, established by UN/CEFACT and OASIS, for constructing a single global electronic market using an open standard based on XML. It provides components for building business documents, for exchanging these documents, and for modeling business processes. Sometimes, ebXML may be regarded as a complex type of Web Services that supports more complex functionalities such as service collaboration. The components of ebXML are published and discovered with the use of ebXML Reg/Rep [1]. On the other hand, Web Services are published and discovered by UDDI [2], a distributed registry standard to provide directory services for Web Services and businesses. Consequently, problems have arisen due to the use of

* This work was supported by the Ministry of Information & Communication, South Korea, under the name of "ITRC Program"

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 163–169, 2005.

different registries, when adopting frameworks, such as that user must likewise use different registries and e-Business systems.

Moreover, there are limitations in the discovery mechanism of these registries[3]. These limitations are caused by the keyword and category-based search features in the registries. The concepts of the Semantic Web vision can be used to lessen the limitations of the semantic expressiveness of the registries. OWL-S[4] is a currently leading Semantic Web Services description language that enriches Web Services descriptions with semantic information from OWL ontologies. To enable the semantic discovery of services, semantic descriptions must be formulated for each service. As such, it is necessary to convert registry entries into OWL-S ontologies to reduce the work volume, and the data inconsistency. To do this, lower OWL-S ontologies were defined to represent a registry information model, mapping relationships between the ontologies and registry entries. Migrated OWL-S ontologies could also be updated periodically using the replication APIs of the registries.

2 Related Works

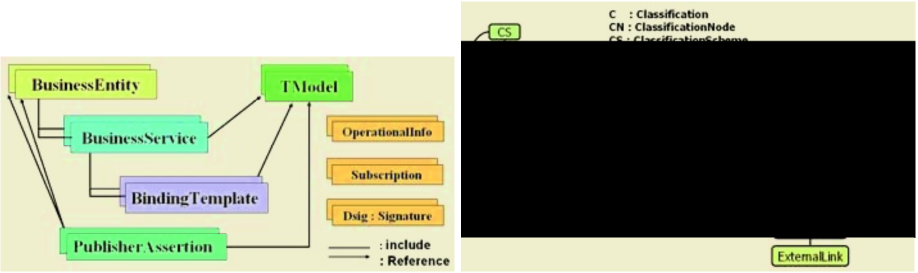
Surveys on the methods of utilizing the UDDI registry for the purpose of addressing ebXML components were made in some studies[5,6]. A way of publishing and discovering Web Services using ebXML Reg/Rep was likewise described in other studies[7, 8]. In [7], a current practice for registering Web Services in an ebXML Reg/Rep without modifying the registry standard was described. In [8], the similarities of both registries were presented, and the UDDI service middleware, which translates UDDI API calls into requests of ebXML registry services using mapping relationships, was developed. In addition, [9] provided a way to uniformly access ebXML 2.0 Reg/Rep and UDDI 2.0 registries using a common data model and common APIs named by Java API for XML Registry (JAXR). JAXR provided a common method of gaming access to the registries using registry drivers, similar to the concept of JDBC DB drivers. The above researches, however, did not seriously support registry integration. [4, 5] only provided for a way to correlate the registries, and [6, 7] only described the use of ebXML Reg/Rep as a substitute for UDDI. JAXR is the only way of integrating registries, but its model and APIs are different from those of the registries. The users who have used JAXR must have had some knowledge about special data models and APIs. Furthermore, the JAXR specification has not updated since 2002, so it does not reflect the current status of registry standards.

The researches that have been conducted with the aim of extending registries so they could support semantic discovery in registry can be classified into two groups, namely: (1) extending legacy Web Services standards by adding semantic annotation to reinforce the discovery function in registries; and (2) a way to preserve semantic advertisements into legacy registries by mapping semantic information into RIM (Registry Information Model). In [10], the advantages of describing service semantics through ontology languages and how to relate the semantics defined with the services advertised in registries were described. In [11, 12], the relationships between WSDL and DAML-S Grounding and the WSDL conversion into DAML-S Grounding, were described. In [13], a way of publishing ontologies into a registry by storing WSDL documents with DAML+OIL semantic annotation as a tModel in UDDI was presented.

In [14], DAML-S Profile was mapped into UDDI RIM and tModels were created to represent non-directly mapped items in DAML-S. In [3], an enhanced UDDI registry capable of storage, matching, and retrieval of semantically service profiles using DAML-S Profile, and a way to correlate DAML-S Profile with registry entries, were presented. In [15], how the various constructs of OWL could be mapped into ebXML classification hierarchies, and how the stored semantics could be queried by using the ebXML query facility, were shown. Although these past studies aimed to generate semantic advertisements for the discovery enhancement of current registries, and to publish the advertisements into registries, they, however, did not suggest how semantic information for services could be generated using legacy registry entries.

3 Comparisons of Information Models

The UDDI information model is composed of five core models and other subsidiary models. *businessEntity*, the root element of UDDI RIM, represents a business partner who creates and provides services. A businessEntity records information such as the name of the business, and its contact information (e.g., address, telephone number, e-mail, the URL of the company web site). A businessEntity is associated with one or more *businessServices*, each of which is a description of the specific services that a business partner provides. In turn, a businessService is associated with one or more *bindingTemplates* that specify the service access points from HTTP, e-mail, and fax to telephone. A *tModel* defines attributes that can be used to specify additional information about the services. UDDI uses two types of tModels: (1) a tModel that expresses the technical specifications of the service, such as the protocols that they adhere to or the interchange formats and (2) a tModel that expresses abstract specifications about the service within predefined classification and identification schemes. *publisherAssertions* are used to allow a registered businessEntity to be linked in a manner that conveys a specific type of relationships. *operationalInfo* includes the date and time that the registry entry was created or modified and the identity of the publishers.



[redacted]

Fig. 1. UDDI and ebXML registry information model

In ebXML RIM, all registry entries are inherited from *RegistryObject*, which provides minimal metadata for registry entries. *Association* is used to define many-to-many relationships between the entries in the RIM. *Service* is a registry entry that provides information on services. *Servicebinding* represents technical information on a specific way to access a specific interfaced offered by a Service entry. *Specifica-*

tionLink provides the linkage between ServiceBinding and one of its technical specifications, which describes how to use the service through ServiceBinding. *ExternalIdentifier* provides additional identifier information on a registry entry such as the DUNS number, the SSN, etc. *ExternalLink* is an entry that contains a named URI of a content that is external to the registry. *User* and *Organization* are entries that are used to provide information about a person or organization that registered other entries. *ClassificationScheme* describes a structured way of classifying or categorizing registry entries. A *ClassificationScheme* defines a tree structure made up of *Classfication-Node*. *Classification* identifies a ClassificationScheme instance and taxonomy value defined within the classification scheme.

UDDI, ebXML RIM, and OWL-S directly use XSD types or those with type inheriting. As shown in Table 1, there is no essential difference among the data types of RIM, and OWL-S. The only difference lies in the object ID generation schemes. While ebXML uses only UUID to identify registry entries, UDDI uses either a UUID or a URI-based key. On the other hand, OWL-S uses rdf:id, which must start with an alphabet or "_".

Table 1. Comparison of datatypes in OWL-S, UDDI, ebXML RIM

ebXML RIM V2.5	**UDDI V 3.0**	**OWL-S**	**XSD Datatype**
Boolean	Boolean	-	Boolean
String, String(4,8,16,21) ShortName, LongName FreeFormText	String(50, 80, 255 ,4096, 8192)	String	String
Id: UUID(Universal Sequence Identifier)	Id: UUID or URI-based key scheme	rdf:id	rdf:id
URI	anyURI	anyURI	anyURI
DateTime	DateTime	-	DateTime

More differences can be found, however, when these models are compared in their entirety. The differences are as follows:

- **Central of registry entries –** In ebXML Reg/Rep, each registry entry is registered independently and the entries are linked with association entries. On the other hand, in UDDI, a businessEntity holds other core registry models into its content model as sub elements except tModels.
- **Business description –** UDDI specifies a business description only with businessEntity, but in ebXML Reg/Rep, the descriptions are specified in conjunction with the User and Organization entries.
- **Object relationships –** While the user can freely specify relationships between entries in OWL-S and ebXML Reg/Rep, UDDI supports only the relationships between two businessEntities using publisherAssertion.
- **URI access to registry entries –** ebXML Reg/Rep provides a REST interface for each registry entry so the user can fetch registry entries separately with UUID via Internet. On the other hand, UDDI specified a discoveryURL that points to Web-addressable (via HTTP Get) discovery documents that contains only a business-Entity and subelements.

- **Category system –** Category systems such as UNSPSC, and NAICS could be built in ebXML Reg/Rep using ClassificationScheme, and ClassificationNode. On the other hand, UDDI, and OWL-S do not support building category systems in their internals. They just keep the codes and values of a category system in the related entries as a string type.
- **Registering external resources –** All external objects that cannot be described using RIM such as WSDL documents and external identification systems are stored using tModels in UDDI. On the other hand, ebXML Reg/Rep describes external objects with ExternalLink, ExtrinsicObject, and ExternalIdentifier according to the type of objects.
- **Information model extensibility –** ebXML Reg/Rep provides a dynamic way of adding arbitrary attributes to each registry entry by associating entries with "slot" entries composed of a collection of name and values. UDDI only provides a way to extend a businessEntity entry using a businessEntityExt, a businessEntity entry that has an XSD AnyType element.

4 Mapping Information Models and System Architecture

Fig. 2 shows how to map UDDI ebXML RIM into OWL-S was achieved in this paper. Most of registry entries that describe businesses and services are directly mapped into OWL-S ServiceProfile, while binding information and WSDL are referred to or converted into OWL-S ServiceGrounding using the algorithms described in [10] and 11].

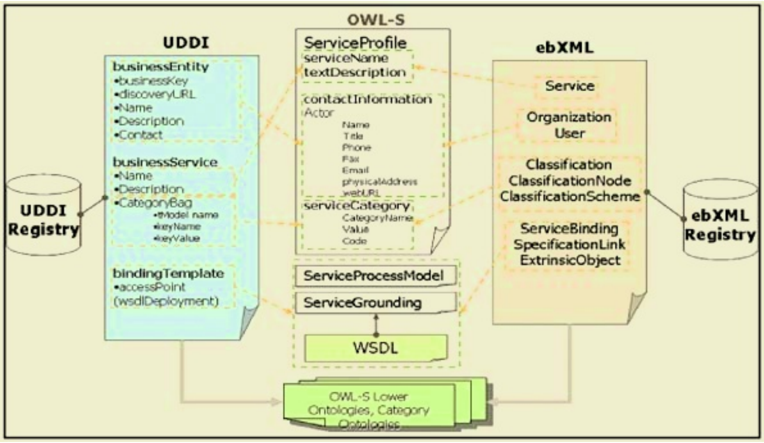


Fig. 2. Mapping UDDI, ebXML RIM into OWL-S Ontology

All entries in registries, however, could not be mapped directly into OWL-S because of the discrepancies described in section 3. To address this problem, OWL-S was extended to specify non-directly mapped entries by creating new OWL classes. Newly created OWL classes were referred to OWL-S classes as object type properties. The additional OWL classes are shown in Fig. 3. First, the OWL-S Actor class was extended to pose *ActorIdentifier* and *ActorCategory* to capture identifier and category values for businesses. Second, a *ServiceBinding* class was created to address the physical service endpoint and other external resources in using the service. Although OWL-S ServiceGrounding supports the signature for service interface, it does not provide the technical information needed to access services such as protocols, security policy, etc. Lastly, a *TechnicalSpecification* class was defined to describe external objects represented by tModels, externalLinks, ExternalIdentifiers in each registry.

Category systems stored in or referred by entries were also converted to OWL classes. With the category concepts represented as OWL classes, the hierarchy of the category system, which is useful for the discovery of services, can be preserved.

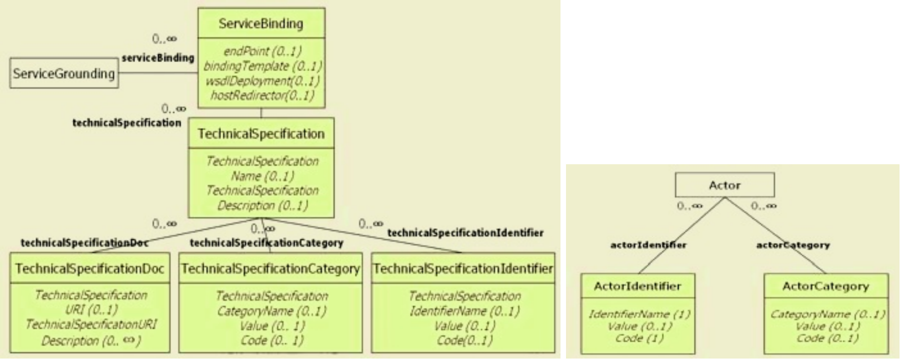

ServiceBinding
endPoint (0..1)
bindingTemplate (0..1)
wsdlDeployment(0..1)
hostRedirector(0..1)
0..∞
serviceBinding
ServiceGrounding
0..∞
technicalSpecification
TechnicalSpecification
TechnicalSpecification Name (0..1)
TechnicalSpecification Description (0..1)
0..∞
technicalSpecificationDoc
0..∞
technicalSpecificationCategory
0..∞
technicalSpecificationIdentifier
TechnicalSpecificationDoc
TechnicalSpecification URI (0..1)
TechnicalSpecificationURI Description (0.. ∞)
TechnicalSpecificationCategory
TechnicalSpecification CategoryName (0..1)
Value (0.. 1)
Code (0.. 1)
TechnicalSpecificationIdentifier
TechnicalSpecification IdentifierName (0..1)
Value (0..1)
Code(0..1)
Actor
0..∞
0..∞
actorIdentifier
actorCategory
ActorIdentifier
IdentifierName (1)
Value (0..1)
Code (1)
ActorCategory
CategoryName (0..1)
Value (0..1)
Code (0..1)

Fig. 3. Additional OWL Classes

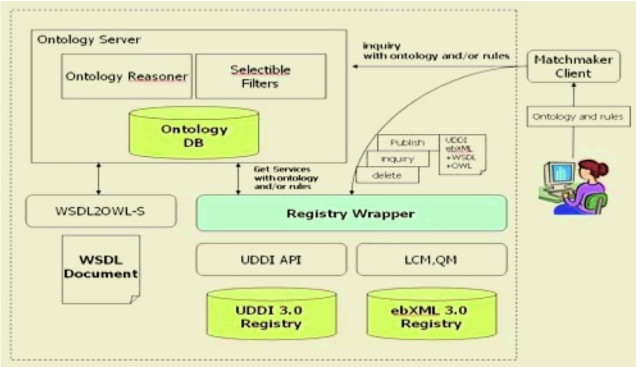


Fig. 4. Registry Integration System

The registry integration system developed in this paper handles the current ebXML RIM V2.5, UDDI V3.0 standard and OWL-S. The UDDI V3.0 registry and the improved FreebXML V2.1 for the purpose of supporting its V2.5 features were used in our implementation. At first, all the entries of each registry are converted into OWL-S ontologies with additional classes. Registry wrapper takes charge of this process and the conversion rules are described with XSLT script + external function set. The converted registry entries are then stored into the ontology DB. For the use of storage, and querying to ontologies, we adopted JENA framework to develop the ontology server. Also, for the purpose of supporting affluent query ability in the ontology server, OWL-QL was used as a query language for the ontologies. Once the registry entries are stored in the ontology server, the server periodically checks the updates of corresponding entries in each registry to maintain data consistency. The server maintained operational information for each ontology, and communicates with the registries using the replication API of each registry.

5 Conclusion

In this paper, we suggested a way to solve not only registry integration but also semantic support problem within registry. Future work directions will comprise investigation for the representation of more various external objects in each registry and business processes such as ebXML CPP, and WSBPEL.

References

1. OASIS ebXML Registry TC.: ebXML Registry Information Model V2.5. http://www.oasis-open.org/committees/regrep/docuemtns/2.5/specs/ebRIM.pdf (2004)
2. OASIS UDDI Specification TC.: UDDI Version 3.0 Specification. http://uddi.rog/pubs/uddi_v3.html (2003)
3. Pokraev, S., Koolwaaij, J., Wibblels, M.: Extending UDDI with context-aware features based on semantic service description. Proc. of ICWS (2003) 184-190
4. The OWL Service Coalition.: OWL-S: Semantic Markup for Web Services. http://www.daml.org/services/owl-s/1.0
5. ebXML Registry Project Team.: Using UDDI to find ebXML Reg/Reps. ebXML White Paper, http://www.ebXML.org/specs/rrUDDI.pdf (2001)
6. OASIS UDDI Spec. TC.: UDDI as the registry for ebXML Component. Technical Note (2002)
7. Chiusano, J.M., Najmi. F.: Registering Web Services in an ebXML Registry. OASIS Registry TC Technical Note, http://xml.coverpages.org/RegisteringWebServices.pdf (2003)
8. Park, J.H., Kim, S.K, et al.: Design and Implementation of a UDDI Service Middleware based on the ebXML Registry. Journal of Korean Information Science Society, Vol 31, Num 3 (2004) 307-319
9. Najmi, F.: Java API for XML Registries Specification (JAXR) 1.0. JSR 98, http://www.jcp.org/en/jsr/detail?id=093 (2002)
10. Dogac, A., Laleci, G., Kabak, Y., Cingil, I.: Exploiting Web Services Semantics: Taxonomies vs. Ontologies. IEEE Data Engineering Bulletin, Vol. 25, No. 4 (2002)
11. Paolucci, M., Srinivasan, N., et al.: Towards a Semantic Choreography of Web Services: From WSDL to DAML-S. Proc. of ICWS (2003) 22-26
12. Martin, D., et al.: Describing Web Services using OWL-S and WSDL. DAML-S Coalition working document, http://www.daml.org/services/owl-s/1.0/owl-swsdl.html (2003)
13. Sivashanmugamm K., et al.: Adding Semantics to Web Services standards. In Proc. of ICWS (2003)
14. Paolucci, M., Kawamura, T., et al.: Importing the Semantic Web in UDDI. In Web Services, E-Business and Semantic Web Workshop, CAiSE (2002) 225-236
15. Dagoac, A., Kabak, Y., Laleci, G.B.: Enriching ebXML Registries with OWL Ontologies for Efficient Service Discovery. 14th Int'l Workshop on Research Issues on Data Engineering: Web Services for e-Commerce and e-Government Application (2004) 69-76

Data Integration Hub for a Hybrid Paper Search

Jungkee Kim[1,2], Geoffrey Fox[2], and Seong Joon Yoo[3]

[1] Department of Computer Science, Florida State University,
Tallahassee FL 32306, USA
jungkkim@cs.fsu.edu
[2] Community Grids Laboratory, Indiana University, Bloomington IN 47404, USA
[3] School of Computer Engineering, Sejong University, 98 Gunja-Dong, Gwangjin-gu,
Seoul, 143-747, Korea

Abstract. In this paper we describe the design of a hybrid search that combines simple metadata search with a traditional keyword search over unstructured context data. This paradigm provides the inquirer additional options to narrow the search with some semantic aspects through the XML metadata query. We demonstrate a paper search for a case study of the hybrid search, and describe a data integration hub to integrate those data dispersed on the Net.

1 Introduction

To discover and share heterogeneous resources on the Net has been a long term challenge since computer communication networks were popularized. There are two traditional approaches to organizing the data to be searched—one is structured data and the other is unstructured data. A Web search engine is a typical example of search on the Internet. Its technologies are rooted in information retrieval that represents search over the unstructured data.

Web search engines provide clues for resource location, but they have no semantic schema and often produce meaningless keyword search results. The Semantic Web is a ambitious extension of the Web. It also includes multiple relation links with directed labeled graphs by which machines like Web crawlers can interpret the relationship between resources. Meanwhile the ordinary Web has a single relationship and a machine cannot infer further meaning. To represent the relations of the object on the Web, the object terms should be defined under a specific description-an ontology. Domain experts are usually needed to design an ontology due to the sophisticated definition required. Currently, many Web pages included no such semantic content, and no unified definition of general semantic agreement exists.

Our hybrid keyword search aims to give an intermediate search paradigm on the Internet—providing semantic value through XML metadata that are simpler than those of the Semantic Web. In this paper we describe our design of hybrid search systems. In earlier experiments, we had suffered from performance problems in a local level, and we proposed scalable hybrid search on distributed environments [4, 5]. In the architecture, a group of independent search

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 170–176, 2005.

providers share their information through their own search systems. But some data providers, who possess small amounts of data, may join such group. They may not want to develop their own search services. Otherwise, participation of many nodes possessing small data in a group will increase the communication traffic and drop a chance to reach the target information under Time-to-Live (TTL) strategy. Partial integration may be one possible method to increase the data portion queried in the search group. In this paper we also present our architecture for data integration hub, which is an application communicating through a message broker with centralized control. This hub can act as a partial integrator on distributed databases or peer-to-peer systems.

This paper is organized as follows. In the next section we describe one of our hybrid keyword search architectures. We present a data integration hub to integrate those papers in Section 3. In Section 4, we summarize and conclude.

2 Hybrid Keyword Search

Our hybrid keyword search combines metadata search with a traditional keyword search over unstructured context data. Each chunk of unstructured data, usually represented by a file, has an assigned metadata. We use XML—the de facto standard format for information exchange between machines—as a metalanguage for the metadata. To demonstrate the practicality of the hybrid keyword search, we design and evaluate hybrid search systems based on a native XML database and a file system based text search library, as well as a market-leading relational database management system that integrates XML and text management.

We have already introduced a relational database based implementation [3, 4]. It utilized an XML-enabled relational Database Management System (DBMS) with nested subqueries to implement the combination of query results against unstructured documents and semistructured metadata. The other implementation is based on a native XML database and a text search library. We use Apache Xindice [2] for XML instance repositories and XML query processing. Jakarta Lucene [1] is used to manage context query over unstructured data in our hybrid search. The query processing architecture is shown in figure 1.

In the Xindice database, we associate an XML instance with an unstructured document by assigning the file name for the document as the key of the XML instance. For example, an XML instance in a file named “pt1.xml” with unstructured data in the file “apaper.pdf” by inserting the XML instance into the data collection as follows:

```
xindice ad -c /a_collection_path -f pt1.xml -n apaper.pdf
```

The assigned key—the file name of unstructured document in the example—as an attribute value on the root element of the XML created by executing an XPath query. For example the query result may start:

```
<bibliography src:col="/a_collection_path" src:key="apaper.pdf"
xmlns:src="http://xml.apache.org/xindice/Query">
```

The key and result XML tuples are stored in a hash table. Another keyword search against unstructured documents returns names of documents which in-

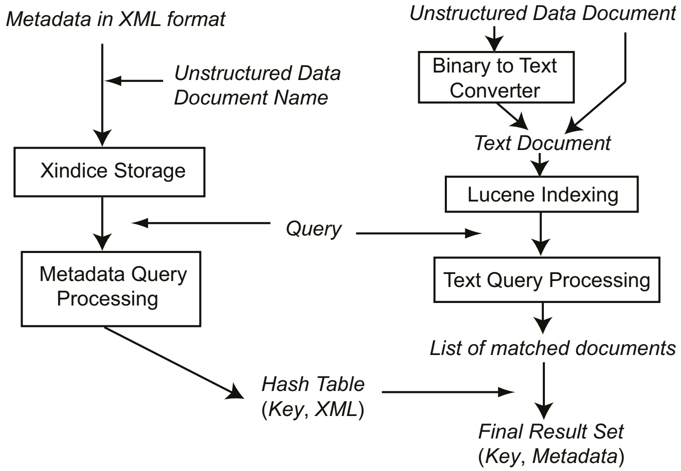

Metadata in XML format
Unstructured Data Document Name
Xindice Storage
Metadata Query Processing
Query
Hash Table (Key, XML)
Unstructured Data Document
Binary to Text Converter
Text Document
Lucene Indexing
Text Query Processing
List of matched documents
Final Result Set (Key, Metadata)

Fig. 1. Query Processing Architecture

clude the given keyword in the text. The returned names are mapped in the hash table and the combined results are collected in a Java hash table.

For efficient text search, Jakarta Lucene provides an index class along with a stop word filtering analyzer class. The analyzer filters out stop words—articles and other words that are meaningless to the search. Some binary format files should be converted to pure text files before indexing. We pass the binary file name as well as the text file name as parameters to the index generating class in order to indicate the original document format.

2.1 Case Study: Hybrid Paper Search

The initial demonstration of the hybrid search is in a simplified model—a paper search. The paper search is a content search across various types of documents. Each document has metadata presented as an XML instance. An example XML instance is shown in figure 2.

In this demonstration we use a relational DBMS instead of a native XML database. Two relational database tables are used for metadata and documents. For the XML instances representing the metadata, the XMLType of Oracle 9i [6] is used. A column with *BFILE* large object type is used for the external document table. Those large object rows are indexed using Oracle Text—a text management system integrated into Oracle DBMS. Through an Oracle Text index, we can search the target content. The document table has a special attribute for a document type—*BINARY* or *TEXT*. This attribute is necessary for the filtering option of Oracle Text. Oracle Text filters binary files to pure text instances before making an index. A one-to-one relationship set is used for relation between the paper document and metadata, but a one-to-many or many-to-many relationship set could be used for other applications. The relationship table is

```
<bibliography>
  <authors>
    <author>J. Kim</author>
    <author>O. Balsoy</author>
    <author>M. Pierce</author>
    <author>G. Fox</author>
  </authors>
  <title>Design of a Hybrid Search in the OKC</title>
  <source>Proceedings of the International Conference on IKS</source>
  <year>2002</year>
</bibliography>
```

Fig. 2. An Example XML Instance

not necessary in one-to-one relationships, but it is essential to decompose relations to avoid anomalies in one-to-many or many-to-many relationship. With two data tables and a relationship table we can query keywords in the content, which can be associated with particular metadata through nested subqueries. For example, we can find documents with a keyword "XML" and published in 2002. Figure 3 shows the relational schema of our hybrid paper search.

papers(paperND: string, descriptions: XMLType)
paperfiles(filename: string, doctype: string, contents: BFILE)
filelocator(paperND: string, filename: string)

Fig. 3. Relational Schema of Hybrid Paper Search

3 Data Integration Hub for Hybrid Search

In our earlier work [5] we assumed query clients only read resource in the other nodes. In this paper we take into account that clients may desire to be a data provider without providing an independent search service. Several or many data providers can share their information through a centralized database. They may upload, query, and download unstructured data with attached metadata via a central DBMS.

Another aspect for the integration hub, which was not introduced in the local hybrid paper search, is the metadata validation against XML schema. The storage for the local hybrid paper search is managed by database administrators, but ordinary users can upload their own data to the database of the integration hub. We utilize an Oracle DB operation to check the validity of metadata presented in XML instances against a registered XML schema. The XML schema for our hybrid paper search is shown in figure 4.

We demonstrate a data integration hub on the central DBMS using message-oriented middleware. This integration hub is an integrated version of the hybrid

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="qualified" attributeFormDefault="unqualified">
  <xs:element name="bibliography">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="authors">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="author" type="xs:string"
                maxOccurs="unbounded"/>
            </xs:sequence>
          </xs:complexType>
        </xs:element>
        <xs:element name="title" type="xs:string"/>
        <xs:element name="source" type="xs:string"/>
        <xs:element name="year" type="xs:decimal"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

Fig. 4. An XML Schema Example for Hybrid Paper Search

paper search introduced in section 2.1. The general architecture of a data integration hub is summarized in figure 5. Those clients can communicate with an integrant through a message broker—NaradaBrokering [7]. This broker includes JMS compliant topic-based communication—a publish/subscribe model. As in the read-only query case, clients are publishers and the integrant is a subscriber. Clients and the integrant use the same topic. The integrant manages uploading of the metadata and unstructured data, query wrapping, and file transfer. Requests are classified by the job property, which is attached to the message sent from client to integrant. A *temporary topic* is also delivered to the integrant. This temporary topic provides unique identification for the requester client, and the integrant can return back to the client by publishing the message on a dynamic virtual channel—a temporary topic whose session object was attached to the message.

We use different message types for each service request, as follows:

– Upload request: initially the client should request to upload to the integrant, before the unstructured data files are sent. The message for this request includes a file name of unstructured data, metadata, and unique client user name. They are string contents in a `MapMessage` type message. The integrant checks the validity of the metadata against XML schema as clients may provide no well-formed or valid data. The central database should avoid redundancy for the unstructured data stored in a file system. The name and directory of the unstructured data, which is generated by combining the user name and the file name, is used for a primary key.

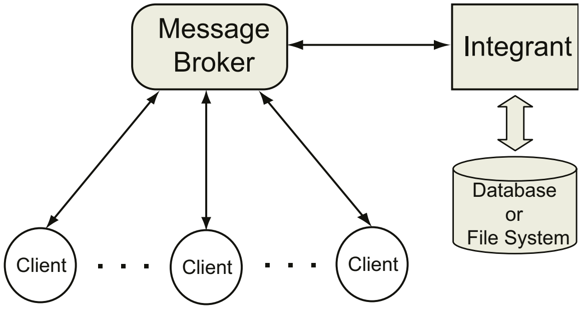


Fig. 5. An Integration Hub Architecture

- File upload: after an upload request is granted by the integrant, the client can send a file for the unstructured data. For the file uploading, we use a `ByteMessage` type message, which is appropriate for sending a stream of uninterpreted bytes. The client sends a header message first, and the file is published buffer by buffer. A job property, a file name, a user name, and a temporary topic are included in the header message. A JMS message has a unique ID when it is created, and the ID is attached to each message. The integrant can classify those file upload messages by extracting the message ID. This mechanism is necessary because several file uploads can occur simultaneously.
- Query request: clients are publishers for a query topic, and the integrant subscribe on the same topic. A `MapMessage` type message, which includes query parameters and properties, is published to the integrant. The query results are delivered back to the inquiry client subscribed to a temporary topic.
- Download request: the target unstructured data in a file can be obtained from the query request. The client subsequently requests a file download with a `MapMessage` type message that includes the file name and a temporary topic. The listener on the client then captures the `ByteMessage` type message published from the integrant, and the target file is written message by message on the client machine. Each message includes the file name property. The message broker is responsible for preserving order of message transfer.

The database schema of our data integration hub are similar to those in the hybrid paper search in figure 3, but there is an additional table for the file uploading for the unstructured data—a temporary file locator table. Our system allows the file upload on a temporary directory only, and moves the files to the designated directory later. When a user request a file upload and incidentally the same file name already exists, a new file name is assigned for the final destination by the integrant. The original file name is stored in the table for metadata, but the actual file name is stored in the unstructured data table. We assume that a user does not assign the same file name to two or more different unstructured data. A naming and directory for each row in the paper metadata table is generated from combining the unique user name and the file name, and it makes the naming and directory to a potential primary key.

Each data integration hub has a message broker and an integrant. A group of data integration hubs may provide a global search over a distributed information system by using the cooperative network features built in to NaradaBrokering, or by using an additional network layer—a peer-to-peer overlay network.

4 Conclusion

In this paper we described our approaches to hybrid search at a local level with a case study of a hybrid paper search. Our demonstration showed the possibility of the hybrid search paradigm for a practical semantic integration. We had another case study of a data integration hub, which is an application communicating through the message broker with a centralized control. The integration hub can be a search service node in a distributed database, or a peer in a peer-to-peer overlay network under more scalable environment [4, 5]. This scalable generalization may have a practical bridging role for information search—providing semantic value through metadata whose implementation are simpler than those of the Semantic Web.

References

1. Apache Software Foundation. Jakarta Lucene.
World Wide Web. http://jakarta.apache.org/lucene/.
2. Apache Software Foundation. Xindice.
World Wide Web. http://xml.apache.org/xindice/.
3. J. Kim, O. Balsoy, M. Pierce, and G. Fox. Design of a Hybrid Search in the Online Knowledge Center. In *Proceedings of the IASTED International Conference on Information and Knowledge Sharing*, November 2002.
4. J. Kim and G. Fox. A Hybrid Keyword Search across Peer-to-Peer Federated Databases. In *Proceedings of East-European Conference on Advances in Databases and Information Systems (ADBIS)*, September 2004.
5. J. Kim and G. Fox. Scalable Hybrid Search on Distributed Databases. In *Proceedings of International Workshop of Autonomic Distributed Data and Storage Systems Management (To appear)*, Lecture Notes in Computer Science. Springer, May 2005. Available at http://www.cs.fsu.edu/˜jungkkim/paper/p2pDDBS.pdf.
6. Oracle Corporation. *Oracle9i Application Developer's Guide—XML*, June 2001.
7. S. Pallickara and G. C. Fox. NaradaBrokering: A Distributed Middleware Framework and Architecture for Enabling Durable Peer-to-Peer Grids. In *Proceedings of International Middleware Conference*, June 2003.

Effective Information Sharing Using Concept Mapping*

Keonsoo Lee[1], Wonil Kim[2,**], and Minkoo Kim[2]

[1] Graduate School of Information and Communication, Ajou University, Suwon, Korea
lks7256@ajou.ac.kr

[2] College of Electronics and Information Engineering, Sejong University, Seoul, Korea
wikim@sejong.ac.kr, minkoo@ajou.ac.kr

Abstract. In this paper, we propose a concept mapping method that can be used to find relations among concepts of different ontology. In order to find the relevant relation, this method uses three kinds of knowledge types such as lexical knowledge, domain knowledge and structure information. When the relations are retrieved, each relation is evaluated by comparing instances of each concept which is involved in that relation. With this method, the concept of one ontology can be mapped to other concept of the other ontology, which enables information sharing.

1 Introduction

It is hard work to find desired information from treasure house provided by IT such as Internet. Information retrieval can be one solution [16]. However, these methods are mainly based on the user's query and usually the result of one engine is isolated from other engine. It means that each search engine should find the similarity between query and document set even if there are results of other search engines. If the ontology, which is the formal specification of a shared conceptualization used by one search engine, is understood by other engines, the searching process of other engines can be concise. When the result of the first engine indicates that the user's intention is 'Tom Cruise's resent movie' which maybe hided from the user's query, the other engine that understands this result can compare the user's query with the document set which is classified by these annotation, such as 'Actor: Tom Cruise, Time: Resent, Subject: Movie'. This is the advantage of using ontology.

As the Fig. 1 shows, ontology can be used in three ways for representing information. The first way is using one global ontology. As there is only one ontology, any application or user who understands the ontology can use all the information without additional process. However, making a global ontology that is acceptable to all the participants is almost impossible. Even if such ontology is made, the size of global ontology must be huge. The second way is using several local ontologies. In this case, the information providers can use their own specified ontology and can represent what they want to show without other's ontology. Therefore, the information which is represented by one specified ontology, should be transformed to be understood by

[*] This research is supported by the ubiquitous Autonomic Computing and Network Project, the Ministry of Information and Communication (MIC) 21st Century Frontier R&D Program in Korea

[**] Author for correspondence: +82-2-3408-3795

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 177–183, 2005.

those who uses other ontology. The last way of using ontology is a hybrid one. Several local ontologies are used and the local ontology references the global ontology. All the local ontologies are inherited the global ontology, the relation among different ontologies can retrieved more easily referring the global ontology. However, as the second way, the mapping solution between different local ontologies is also necessary, which may be simply than that the second way employs.

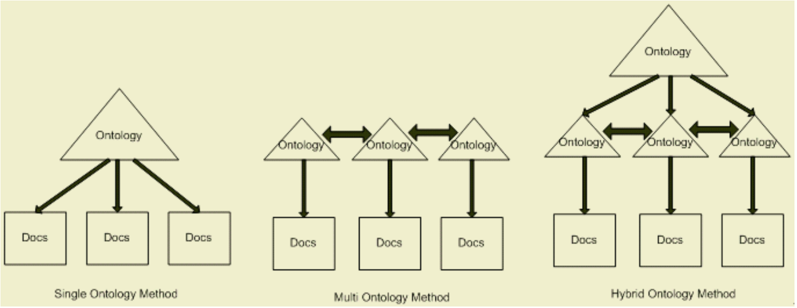


Fig. 1. Information representation method using Ontology

The ontology mapping method that is proposed in this paper is for the second and third cases where multiple ontologies are employed. With this method, even the information represented by different ontology can be shared. Related researches are discussed in chapter 2. The proposed method is explained in chapter 3. The simulation of this method and its result is shown in chapter 4. Finally chapter 5 concludes.

2 Related Research

In order to share the information, the applications should understand the meta-information that explains the information. When the ontology is employed for representing this meta-information, all the application can share the information by just understanding the ontology. In 1995, 'Dublin Core' was suggested as meta-data standard [11]. Even though 'Dublin Core' is widely used, it is for bibliographic and its number of elements is just 15. It shows the generality as global ontology has limited the ability of representing the detail of information.

In order to represent the detail characters of specific domain, the generality should be abandoned. The SWEET (Semantic Web for Earth and Environmental Terminology) ontology suggested by JPL[12] and Global Change Master Directory and ESMF (Earth System Modeling Framework) are different onotoloies for the same domain; geography. Even though these ontologies can represent detail of the domain, the information represented by SWEET can not be used by any applications which use ESMF. The SWEET offers the mapping table to other ontologies, which shows the 'Wind Speed' in SWEET is mapped to 'Atmosphere' in ESMF. However, if there are more ontologies, the size of mapping table will be huge. Using mapping table is proper with a few ontologies to be mapped.

Therefore, more dynamic method is necessary such as concepts comparing. In order to find the relation between concepts in different ontologies, additional knowledge needs to be used. Lexical knowledge, domain knowledge or structure information can be the one. The lexical knowledge is about the name of concepts. For example, a concept named ‘Character’ can be assumed to have semantic relation with the concept named ‘Actors’. With the lexical knowledge, the direct relation such as synonym can be inferred. The domain knowledge is about the sense of concept. Let me assume that there are two concepts ‘sound’ and ‘audio’. Without the domain knowledge, the two concepts may be assumed to be equal. But If the ‘sound’ has IS-A relation with ‘music’ and the ‘audio’ has IS-A relation with ‘household appliances’, there is not ‘equal’ relation. The domain knowledge is used to select the sense of concept. The structure information is also used for ontology mapping. If there are two concepts which have the similar attributes such as data type, range, cardinality and domain, they may have some relation. The structure means not only the format of concept but also the whole hierarchy of ontology. These knowledge types are widely used for finding relations between concepts. CtxMATCH [1], CUPID [2] and GLUE [3] are the most famous methods using these knowledge types.

On the other hand, the information represented by the ontology was concerned for ontology mapping. Comparing the information, the concepts that used to represent the same information can be regarded as equivalence [4]. If the same information is represented by two different concepts, there must be a relation between these two concepts.

3 Suggested Method for Ontology Mapping

With ontology, the applications which know the ontology can share the information. This paper proposed an ontology mapping method of information sharing for applications which use different ontologies. Fig. 2 shows the flowchart of propose method.

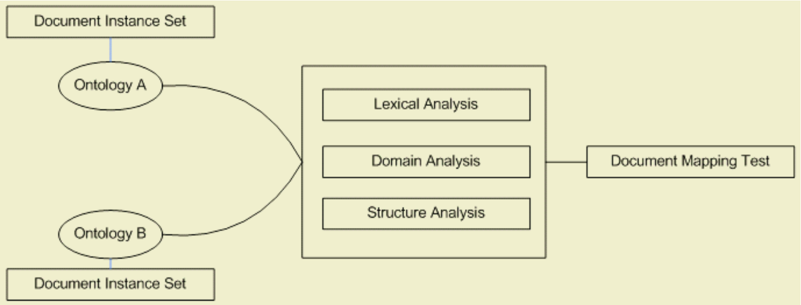


Fig. 2. Flowchart of suggested mapping method

This method consists of 4 steps, lexical analysis, domain analysis, structure analysis and document mapping test. When two ontologies and each ontology’s instance set are inserted, the lexical analysis process is executed. In this step, all the concepts are analyzed according to its lexical information. The synonyms and antonyms are retrieved for all the sense. In domain analysis step, the relevant sense is selected. Let me assume that there is a concept named ‘audio’. This concept has three senses. First is a sound as the audible part of a transmitted signal. Second is an audio frequency as an audible acoustic wave frequency. The third is an audio as the sound elements of television. This is the result of the lexical analysis. Then, according to the domain of the ontology, the true sense of audio is selected. If the domain of this ontology where this 'audio' concept is involved, is 'household appliances', this concept has the third sense and the synonym list has only one element, 'audio'. All of the concepts in the ontology are assigned with their own synonym list and comparing this list, the 'Equivalence' concept relation is endowed. Table 1 shows the list of concept relations. Every two concept pair has one of the relations in table 1. The 'Equivalence' concept relation is assigned when the two concepts have the most similar synonym list. Once the 'Equivalence' concept relation is assigned, other relations are endowed through the structure analysis step. If concept A and concept B have 'Equivalence' concept relation, concept A has 'Less general than' relation with parents concepts of concept B. Likewise, all of the parents concepts of concept B has 'More general than' relation with child concepts of concept A.

Table 1. Concept relation

Relation	*Meaning*
Equivalence	Concepts are equal.
Less general than	Former is a sub notion of latter.
More general than	Latter is a sub notion of former.
Not related	Two concepts are not related.

When the 3rd step, structure analysis is over, all of the concepts have relations for each concept. With the relations of each concept to all the concepts in the other ontology, the concept relation metric table is structured. Each relation in this metric table is verified by document mapping test, the 4th step. This verification is tested by classifying the concepts instance. If concept 1 and concept A have 'Equivalence' concept relation, the instances of concept 1 are classified by concept A. If the result exceeds the threshold, the relation is verified. Through the 4th step, the relation metric table is weeded out and new table with relevant relations is made. With this table, two different ontologies can be mapped and the application which does not know the other ontology, can understand the information which represented by the other ontology. As this method uses three types of knowledge, it can match the concepts with independency of environment such as whether the instances of classes are shared or not, whether the descriptions of classes are provided or not.

4 Simulation and Result

In order to simulate the method proposed in this paper, we postulate these conditions. As shown in Fig. 3, there are two different ontologies. These ontoloies have their own instance set. These instance set are used for document mapping test, 4th step. We constructed these ontologies and their instance set from the web portal, 'www.yahoo.com' and 'www.altavista.com'. The whole directory information is so huge that we selected sub category, 'Multi-Media' in 'Movie'.

For the lexical analysis, 1st step, we used the dictionary provided by WordNet [10]. The name of each concept is sent to WordNet and the WordNet retrieves the list of senses the name has. Then, the domain analysis, 2nd step, is carried out. The human being selects the relevant sense from the list. From the selected sense, a synonym list is assigned to each concept. From these synonym lists, the concept which has the most similar list is assumed to have the ‘Equivalence’ relation. With this ‘Equivalence’ relation and the structure information, the ‘More general than’, ‘Less general than’ and ‘Not Related’ relations are assigned, 3rd step. If the ‘video clips’ concept of yahoo has ‘Equivalence’ relation with ‘video’ concept of altavista, the ‘video clips’ concept has “More general than’ relation with ‘short films’ concept which is the child concept of ‘video’. From these processes, the concept relations are made.

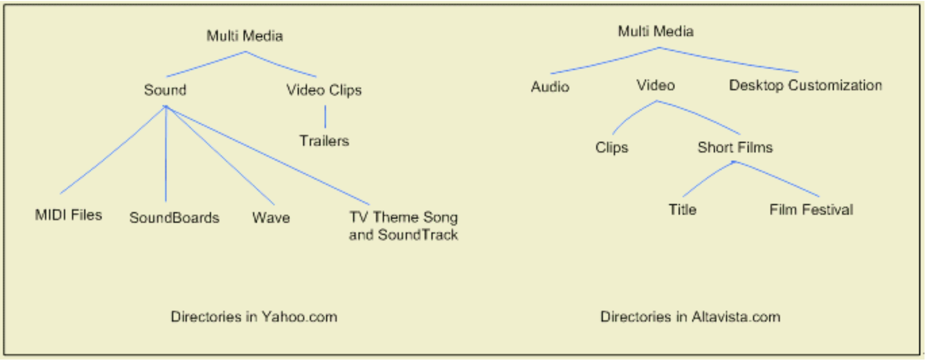


Fig. 3. Ontology example

Each relation between concepts is testified through document mapping test, 4th step. For this step, we selected ten documents for each concept. Then, we checked whether the documents classified for the concept can be classified to the concepts of the other ontology, assigned ‘Equivalence’ relation. The checking method is simplified TF-IDF one. If the frequency of original concept in the document and the frequency of mapped concept are similar, the difference is less than 15%, we assumed the relation is relevant. For example, the ‘sound’ concept in yahoo is assumed to have equivalence relation with ‘audio’ concept in altavista. Therefore, the mapping test carried out to check whether this relation is relevant or not. First, ten documents are selected from the yahoo and the frequency of ‘sound’ and ‘audio’ are counted. The 8 out of 10 documents have similar frequency number and we concluded the relation ‘Equivalence’ is relevant. This process is carried out for every relation and the final result of ‘Equivalence’ relation can be seen in Fig. 4.

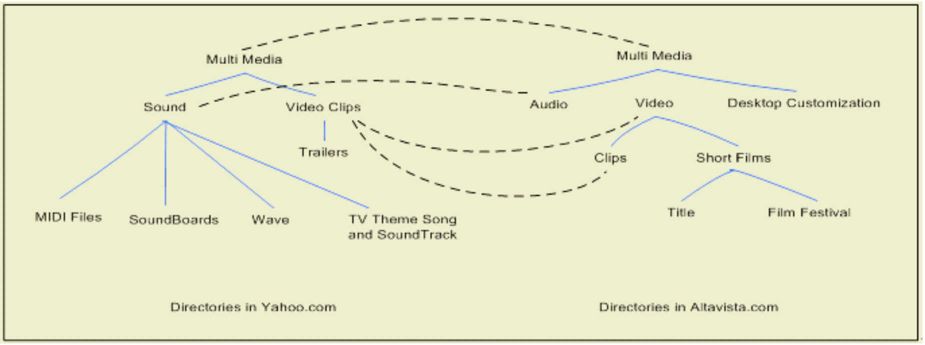
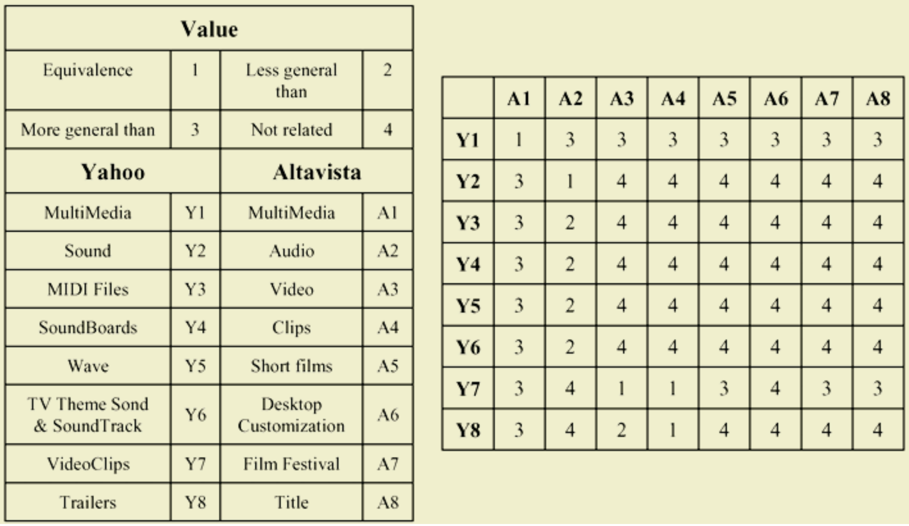


Fig. 4. Result of ontology mapping

Table 2 shows the final metric of concept relation. Most of the relations such as ‘Equivalence’, ‘More general than’ and ‘Less general than’ pass the test of comparing information step. The ‘Not Related’ relation is removed for the concept never appears in the documents as the meaning of the relation.

Table 2. Concept relation metric table



Value			
Equivalence	1	Less general than	2
More general than	3	Not related	4
Yahoo		**Altavista**	
MultiMedia	Y1	MultiMedia	A1
Sound	Y2	Audio	A2
MIDI Files	Y3	Video	A3
SoundBoards	Y4	Clips	A4
Wave	Y5	Short films	A5
TV Theme Sond & SoundTrack	Y6	Desktop Customization	A6
VideoClips	Y7	Film Festival	A7
Trailers	Y8	Title	A8

	A1	A2	A3	A4	A5	A6	A7	A8
Y1	1	3	3	3	3	3	3	3
Y2	3	1	4	4	4	4	4	4
Y3	3	2	4	4	4	4	4	4
Y4	3	2	4	4	4	4	4	4
Y5	3	2	4	4	4	4	4	4
Y6	3	2	4	4	4	4	4	4
Y7	3	4	1	1	3	4	3	3
Y8	3	4	2	1	4	4	4	4

5 Conclusion

This paper proposes a mapping method between concepts in different ontologies. In order to find relevant relations, lexical knowledge, domain knowledge and structure knowledge are used. The retrieved relations are testified through document classification. In order to use lexical information, WordNet is used. WordNet returns the sense list of given word; the name of concept. From this sense list, the most relevant sense is selected by domain knowledge. With this process, the synonym list for each concept retrieved. Then comparing these lists, the concept relations are made. We used 4 relation types, such as ‘Equivalence’, ‘More general than’, ‘Less general than’ and ‘Not Related’. The ‘Equivalence’ is assigned to the concept which has the most similar synonym list. And ‘More general than’, and ‘Less general than’ are assigned to the concepts according to the structure of ontology. If concept A is a child concept of concept B and equivalent concept 1, the concept B is more general than concept 1. At the same time the concept 1 is less general than concept B. These relations are testified by information classification. As the result of test shows, this method proposed in this paper can find the relevant relations. With this method, applications can use any information represented even by unfamiliar ontology.

References

1. Paolo B., Luciano S., Stefano Z.: Semantic Coordination: A New Approach and an Application. In the Semantic Web-ISWC2003 page 130-145.
2. Jayant M., Philip A.B., Erhard R.: Generic schema matching with cupid. In The VLDB Journal, pages 49-58, 2001
3. Doan, J.M., Domingos, P., Halevy, A.: Learning to map between ontologies on the semantic web. 11th International WWW Conference, Hawaii, 2002.
4. Xiaomeng Su.: A Text Categorization Perspective for Ontology Mapping. Technical report, Department of Computer and Informatin Science, Norwegian University of Science and Technology, 2002.
5. Klein, M.: Combining and relating ontologies: an analysis of problems and solutions. In proceedings of the 17th International Joint Conference on Artificial Intelligence, Workshop: Ontologies and Information Sharing, Seattle, USA, 2001
6. Rahm E., Bernstein P. A.: A survey of approaches to automatic schema matching. VLDB journal 10(4): 334-350, 2001
7. Noy N., Musen M.: PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. Proceeding of the AAAI-00 Conference. Austin, USA
8. Maedche, S. Staab.: Ontology learning for the semantic web. IEEE Intelligent Systems, 16(2):72-79, March/April 2001
9. M. S. Lacher, G. Groh.: Facilitating the exchange of explicit knowledge through ontology mappings. In the 14th International FLAIRS Conference, Key West, FL, 2001. AAAI Press
10. WordNet: http://www.cogsci.princeton.edu/~wn
11. Dublin Core: http://dublincore.org
12. SWEET Ontology: http://sweet.jpl.nasa.gov/ontology
13. Protégé: http://protege.stanford.edu
14. Recardo Baeza-Yates, Berthier Ribeiro-Neto,: Modern Information Retrieval, ACM Press, Addison-Wesley ISBN: 0-201-39829

Ontology Supported Semantic Simplification of Large Data Sets of Industrial Plant CAD Models for Design Review Visualization

Jorge Posada[1], Carlos Toro[1], Stefan Wundrak[2], and André Stork[2]

[1] VICOMtech Research Centre, Mikeletegi Pasealekua 57, 20009 Donostia, Spain
{jposada,ctoro}@vicomtech.es
http://www.vicomtech.es

[2] Fraunhofer Institute for Computer Graphics, Franhofer Strasse 5, 20009 Darmstadt, Germany
{andre.stork,stefan.wundrak}@igd.fhg.de
http://www.igd.fhg.de/igd-a2

Abstract. We present in this article a semantic compression system for design review visualization of large data sets in the domain of industrial plant design. This system introduces semantic aspects that improve traditional Computer Graphics techniques used for interactive walkthroughs. We complement previous works introducing new modules and algorithms for the automatic categorization, simplification and user-oriented adaptation of engineering components in the model, and base directly our work on a full Ontology based on international standards for product data in this domain (ISO-STEP 10303-AP227).

1 Introduction

Large Model Visualization (LMV) for industrial plant review is a well-researched area in Computer Graphics. ([3], [6], [8]). The possibility to have interactive walkthroughs for very large geometric datasets offers clear benefits as it reduces design times while in the meantime helps in the early detection of potential construction problems. The main approaches to the LMV problem presented in the literature are mainly related to algorithms and compression methods to be applied to the geometric entities composing the CAD model ([6], [10]). However no special attention is given to the fact that these models contain well-known engineering parts arranged in concrete shapes. A consequence is to have models of millions (or even billions) of triangles when converted from CAD to VR environments. Redundancy could be more intelligently exploited, taking into account the nature of the models and the semantics contained in the CAD. Many legacy systems, as well as only 3D CAD representations of integrated Plant Information Managers (PIM) systems, are often the only basis for Design Review walkthroughs. Unfortunately, high-level semantics are not fully exploited for visualization in CAD although some semantic information is embedded in the geometry. We present in this article a system for design review visualization of large data sets in the domain of Plant Design, based in the categorization, simplification and semantic compression techniques for the engineering parts in the model (figure 1). We base our work on international standards for product data in this domain -ISO-STEP 10303-AP227 [4], adding semantic criteria to the simplification techniques. This paper is presented as follows, in section 2 some background is introduced. Chapter 3 explains our proposed semantic compression technique. Chapter 4

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 184–190, 2005.

presents a case study with some statistics and results. In Chapter 5 we formulate conclusions.

2 Background

There are 4 families of techniques used commonly in interactive walkthroughs of large databases [6]: (*i*) rendering acceleration, (*ii*) database management, (iii) interactive collision detection, and (*iv*) system integration. As seen in [3] the main acceleration techniques used are visibility culling, object simplification and image-based representations. Geometric simplification techniques e.g. Levels of Detail (LOD), Hierarchical Levels of Detail (HLOD) [5] give good results in handling massive data sets; the integration of LOD and good occlusion culling techniques are usually the key factors to achieve interactive rates in walkthrough systems [1]. On the other hand, there are emerging commercial applications (e.g. *NavisWorks*, *Mantra4D*) that incorporate the latest graphics hardware accelerations as well as many of the classical culling and simplification techniques with good results. We take as a starting point any proprietary geometric 3D CAD representation of an industrial plant. We deliberately assume that no other information is available (e.g. from PIM systems). Then we reconstruct automatically the families of engineering parts in the model, associate those families to the standard, introducing both geometric and semantic object simplification techniques, and present the adapted plant model in an interactive system for design review walkthroughs. The work of this paper is strongly based on our previous work [8]. However, [8] still left open the following topics: *(i)* The standards adaptation module *(ii)* An interaction with the Ontology model.

3 Semantic Related Modules in the Walkthrough Architecture

3.1 The Catalog Reconstruction Module

The 3D CAD model creation in the domain of Plant Design is based in the parametric definition and selection of appropriate engineering parts from specific catalogues. However, the resulting CAD models usually do not contain any explicit instancing information, so the first step towards an increased semantic representation of the model is to group these parts using a matching algorithm [9]. This module traverses the 3D CAD model identifying groups of geometric primitives (we call these groups/families *cells*) automatically, and categorizes them based on geometric similarities.

3.1.1 Searching and Classifying Instances

General methods for searching repeating structures in unorganized sets of geometric primitives exist, but are usually slow on large models [2]. The estimated runtime for models common in LMV may easily exceed a full day. We have developed a fast algorithm for finding instances (repeated cells not sorted, as in a soup of elements no matter their orientation or position in space). No assumption is made regarding the internal order of the primitives inside a cell. This Algorithm is explained in detail in [9].

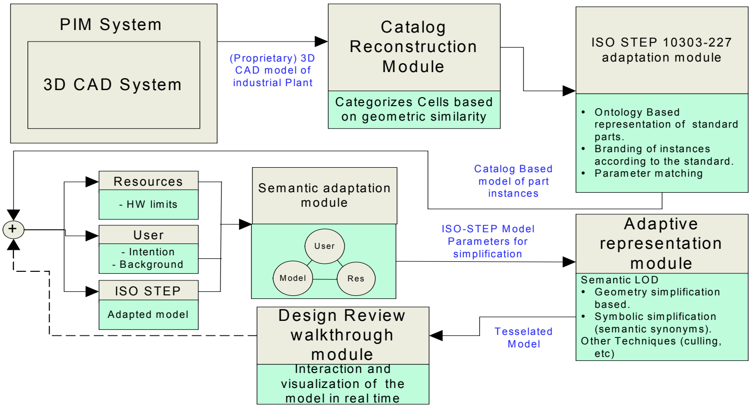

PIM System
3D CAD System
(Proprietary) 3D CAD model of industrial Plant
Catalog Reconstruction Module
Categorizes Cells based on geometric similarity
ISO STEP 10303-227 adaptation module
• Ontology Based representation of standard parts.
• Branding of instances according to the standard.
• Parameter matching
Catalog Based model of part instances
Resources
- HW limits
User
- Intention
- Background
ISO STEP
Adapted model
+
Semantic adaptation module
User
Model
Res
ISO-STEP Model Parameters for simplification
Adaptive representation module
Semantic LOD
• Geometry simplification based.
• Symbolic simplification (semantic synonyms).
Other Techniques (culling, etc)
Tesselated Model
Design Review walkthrough module
Interaction and visualization of the model in real time

Fig. 1. Our semantic walkthrough architecture

3.2 ISO-STEP 10303 Adaptation Module

A 3D model of an Industrial Plant typically has representations of pre-defined engineering parts. These elements are described by an ISO standard (STEP-10303-227 [4]) in the domain of Plant Design. We integrated a module to explicitly associate this semantics to the geometric parts from the reconstruction described in section 3.1.

3.2.1 The ISO STEP 10303-227 Standard

ISO STEP-10303 [4] is the international Standard for the computer interpretable representation and exchange of product data. The objective of STEP is to provide a neutral mechanism capable of describing product data throughout the life cycle independently from any particular system. The nature of this description makes it suitable not only for neutral file exchange, but also as a basis for implementing and sharing product databases and archiving. STEP is a collection of conceptual models. The Application Protocol 227 describes the specifics for plant spatial configuration

3.2.2 Motivation for an STEP Based Ontology Support

We have modeled a full Ontology based on the ISO-STEP standard because our ultimate objective is to have a system where the concepts and relationships of the domain could be modeled and queried using semantic criteria [7], beyond the mere data modeling structures of the norm. This Ontology modeling also allows a more transparent interrogation of the user task/profile that are also modelled as Ontologies.

3.2.3 Construction of the Ontologies

The Ontologies are modelled using Protégé 2000 [9], adapting the tags and relationships (to be more suitable for a knowledge representation model) presented in the ISO STEP standard [4]. This serves as an important contribution to the model part of the semantic triangle described in [7]. The current Ontology of the domain model has a total of 298 classes, 143 slots and 451 frames, and currently represents the 60% of the STEP Application Protocol 227. For the User and Task parts of the semantic triangle, we based our implementation in similar concepts developed by our group in the European Project WIDE (IST-2001-34417).

3.2.4 Interaction with the STEP-Based Ontology

By giving a user task/profile (manager, engineer, etc) the available computer resources and the model (three Ontologies) we query them in order to select an adaptive representation of the model. The model Ontology is filled with the real parameters of the CAD model, and then a semantic association followed by a semantic adaptation allows the visualization enhancement by producing an output that has embedded juts the needed information for each user/task profile and available computer resources.

3.2.5 Semantic Association of Parts with the Standard

In order to add the semantic information we follow a two stages approach (*i*) Name each group of cells after an ISO – STEP compliant concept. We call this process "*Branding*". The user visualizes one representative part of the cell group and matches it with a concept of the Ontology in a graphical concept tree. (*ii)* Once the cell group is associated with a concept in the Ontology domain, the user matches semi-automatically the cell parameters (geometric features) with those parameters specified in the ISO-STEP standard. (We call this process "*Matching*").

3.3 The Semantic Adaptation Module

This module takes as input the adapted 3D CAD model in which the families of cells identified in the Catalog Reconstruction module already correspond to ISO-STEP 10303-227 parts. As explained in ([7],[8],[9]), we have defined a framework in which three factors influence the final adaptation of a 3D CAD model for Design Review walkthroughs: (i) *The user intention and background*, (ii) *the available resources*, and (iii) *the model characteristics*. We introduce explicitly the concepts of *user profile* and *user task*, which influence the final output model in this semantic adaptation module. Thus, the parameters used by different Computer Graphics techniques (such as LOD, culling, etc.) inside the Adaptive Representation module are defined (with a rule-based approach) according to the user needs. In a similar way we take into considerations the available resources (e.g. clusters of PCs vs. single PC, available RAM, etc.) to prepare the walkthrough experience, creating different adapted representations in each case. For example, in Figure 2, in the left side the profile *piping engineer* and the task *piping fixation* forces the system to keep the small clamps and simplifies the elbows, whereas in the right part the profile *manager* and the task *presentation to customers* applies a drop culling technique to the clamps and shows the elbows in geometric detail. We are now moving from the rule-based adaptation system of *user profile* and *available resources* towards a deeper integration with the *model characteristics,* by modelling those two aspects in especial Ontologies that can be integrated with the model Ontology described in 3.2 For a better inference of the right parameters and techniques to use.

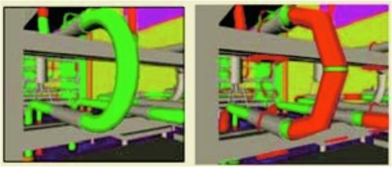


Fig. 2. Semantic adaptation

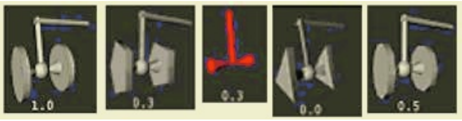
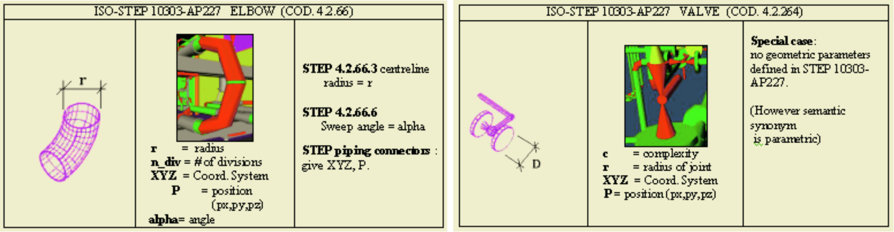


Fig. 3. Geometric LODs vs. Semantic Symbols

3.4 The Adaptative Representation Module

This module receives the adapted model, as well as the parameters for graphical optimization of the final tessellated model displayed in the walkthrough viewer. We have implemented several techniques, although in this article we focus in a special use of the LOD technique that has reported substantial improvement in the walkthrough performance. As explained in section 2, LOD techniques are based on a varying accuracy in the representation of a 3D object. Usually LODs are either automatically generated from the geometric definition of the object or they are modelled adhoc. In both cases the geometric similarity between the LOD and the object is preserved as much as possible. In our approach, once we have the ISO-STEP adapted model, we generate alternative representations according to the parameters given by the previous module: *(i)* We use parametric *geometric LOD* for those components of the model that have the largest influence in the number of triangles generated. These geometric LOD are based on the standard parametric parts of the ISO-STEP 10303-AP227 standard (instead of basing the LODs on the original 3D objects in the CAD model). *(ii) W*e generate in parallel alternative 3D semantic symbols for all components (e.g. 3D symbol for valve) which gives a much higher semantic compression ratio (better compression) without semantic loss for special user profiles and tasks. This of course depends on specific configuration of users/tasks, models and resources. Figure 3 shows an example of the advantage of semantic symbols instead of pure geometric LODs. The Grey objects are geometric LODs generated automatically, whereas the red object is the symbol we used to replace the original elements. At a tessellation complexity of 0.3 this symbol can be rendered with 100 triangles instead of 1000 triangles that were needed for the original. In Figure 4 some ISO-STEP elements are selected to show the adapted representation and the elements to be matched (branding and matching as explained in section 3.2.3).



ISO-STEP 10303-AP227 ELBOW (COD. 4.2.66)

r = radius
n_div = # of divisions
XYZ = Coord. System
P = position (px,py,pz)
alpha= angle

STEP 4.2.66.3 centreline radius = r

STEP 4.2.66.6 Sweep angle = alpha

STEP piping connectors : give XYZ, P.

ISO-STEP 10303-AP227 VALVE (COD. 4.2.264)

c = complexity
r = radius of joint
XYZ = Coord. System
P = position (px,py,pz)

Special case: no geometric parameters defined in STEP 10303-AP227.

(However semantic synonym is parametric)

Fig. 4. Adapted representation of some components: geometric LODs and semantic symbols

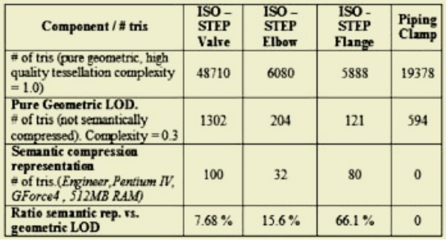
3.5 Design Review Walkthrough Module

Once the semantic data is added and used to simplify the elements via the semantic synonyms, the elements are ready for visualization and walkthrough evaluation. In the Design Review Walkthrough module, we implemented not only the traditional LMV techniques presented in part 2, but also the semantic compression module presented in this article.

4 Case Study – A Chemical Plant, Statistics and Results

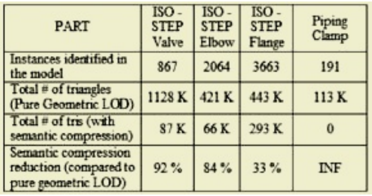
We present in this chapter the results of using our framework in a real-world chemical plant model. The model consist in a large three-story building whose halls are filled with a complex piping system with a lot of curved elements that are very costly to render. After a thorough analysis of the model we found out that more than 65% of the triangles, were produced in the piping system substructure. We have therefore concentrated our efforts in this subsystem and its typical components. With regard to the Catalog Reconstruction Module, it is interesting to see how the elements in this concrete model were grouped. We have found that a high proportion of the total of primitive elements in the model are indeed grouped in cells (65%). This accounts also for a high proportion of the total number of triangles rendered (about 87% of the triangles, even using geometric LOD with complexity = 0.3). From the geometry not organized into cells, another 10% of the triangles come from about 100 complex objects -boilers and tanks- and 3% of the triangles are part of other repeating element like columns, windows, square pipes, etc. The Catalog Reconstruction Module (3.2.) was able to classify the 13147 cells in 1104 families with the Cell Matching algorithm (Table 1). This means that the ISO-STEP 10303-AP227 Adaptation Module (3.3.) was able to classify 82% of the total cell families and relate them to the standard. Table 2 shows the effect of the Semantic Adaptation Module and the Adaptive Representation Module per part type. The results shown are for a piping engineer as user, with check connections task. It is evident from the table that a brute-force, blind conversion with very high quality from the original CAD geometry would create an untraceable model in the practice for design review.

Table 1. Effect of semantic compression for some parts in the Chemical Plant model



Component / # tris	ISO – STEP Valve	ISO – STEP Elbow	ISO - STEP Flange	Piping Clamp
# of tris (pure geometric, high quality tessellation complexity = 1.0)	48710	6080	5888	19378
Pure Geometric LOD. # of tris (not semantically compressed). Complexity = 0.3	1302	204	121	594
Semantic compression representation # of tris.(*Engineer, Pentium IV, GForce4 , 512MB RAM*)	100	32	80	0
Ratio semantic rep. vs. geometric LOD	7.68 %	15.6 %	66.1 %	0

Table 2. Semantic compression reduction



PART	ISO - STEP Valve	ISO - STEP Elbow	ISO - STEP Flange	Piping Clamp
Instances identified in the model	867	2064	3663	191
Total # of triangles (Pure Geometric LOD)	1128 K	421 K	443 K	113 K
Total # of tris (with semantic compression)	87 K	66 K	293 K	0
Semantic compression reduction (compared to pure geometric LOD)	92 %	84 %	33 %	INF

Therefore we take as basis for our comparisons a model already including several simplifications, especially the use of geometric LOD on the original CAD geometry with a complexity of 0.3. This complexity factor in our system is a parameter between 0.0 and 1.0, where 1.0 is the highest accuracy representation. We estimate that a value under 0.3 would create distortions on the tessellated model too evident for the user. The element with the highest reduction (valve), for example, is represented semantically with just 7,68% of the best geometric LOD simplified object. In the case of the clamps, however, the semantic criterion gives an even better hint: the clamps are just *not shown* (drop culling) for this specific task and user. The semantic compression improves in several cases more than 80%-90% from the purely geometric simplification approach, and this especially in those components with highest weight in the tessellated model,which using only geometric LOD plus some culling gave an average number of triangles of 3450 Ktris, with a complexity of 0.3 (already a very good simplification factor). However, applying the semantic compression, we reduced the model in additional 1659 Ktris, for a net reduction of 51% in the total number of triangles between the semantically compressed model with respect to the geometric LOD simplified model.

5 Conclusions and Acknowledgements

We have presented a semantic compression system for design review interactive walkthroughs in the Plant Design domain. The use of the semantics implicit in the geometric model of the and in the user intention and background, have given a sensible improvement in the application of standard computer graphic techniques. In this article we focused mainly in the influence on LODs, improving previous works with new modules and algorithms for automatic categorization, simplification, semantic compression and walkthrough adaptation of a complex plant. In order to achieve generality, we founded our work in the use of Ontologies using international standards (ISO-STEP 10303-227). This work has been partially supported by the European Network of Excellence AIM@SHAPE, IST project 506766 and the Basque Government under the INTEK program.

References

1. Andujar C; Saona-Vazquez C;. Navazo I;. Brunet P.: Integrating occlusion culling and levels of detail through hardly-visibly sets. In Proceedings of Eurographics, 2000.
2. Besl, P.J., Mckay, N.D.: A method for registration of 3D shapes. IEEE Pattern and Machine Intelligence, 14 (2), pp239-256, 1992
3. Funkhouser T.A., Khorramabadi, D., Sequin, C.H., Teller, S.: The UCB system for interactive visualization of large architectural models. Presence, 5(1):13–44, 1996.
4. International Standard ISO 10303, Industrial Application Systems and Integration – Product Data Representation, Application protocol 227, Plant Spatial Configuration., First edition, Geneva, Switzerland 2001
5. Luebke, D, Reddy, M., and others: Level of Detail for 3D Graphics: Computer Graphics and Geometric Modeling. ISBN 1-55860-838-9. 2002.
6. Manocha ,D.: Interactive Walkthroughs of Large Geometric Datasets. In Proc. of Siggraph, USA, 2000.
7. Posada, J; Larzabal, A; Stork, A: Semantic-Based Parametric Control of CAD Model Conversion for Large Model. Cruz-Neira, Virtual Concept 2002. Proceedings. 2002, pp. 38-43.
8. Posada, J; Wundrak, S; Stork, A.; Toro, C: Semantically controlled LMV techniques for plant design review, Proceedings of DETC/CIE 2004 ASME 2004, Salt Lake City, Utah
9. Posada, J; Toro, C, Wundrak, S; Stork, A, "Using Ontologies in the visualization and design review of LMV models" FOMI 2005" Formal Ontologies Meet Industry, Accepted for publication.
10. Shikhare, D., Bhakar, S., Mudur S. P.: Compression of Large 3D Engineering Models using Automatic Discovery of Repeating Geometric Features. VMV 2001 Stuttgart, Germany, November 2001.

EISCO: Enterprise Information System Contextual Ontologies Project

Rami Rifaieh and Nabila Aïcha Benharkat

National Institute of Applied Sciences, LIRIS-INSA de Lyon,
Bât B.Pascal, 7, Avenue J.Capelle, 69621 Villeurbanne, France
{rami.rifaieh,nabila.benharkat}@insa-lyon.fr

Abstract. The enterprise information systems (EIS) offer the cornerstone for managing enterprise business, applying strategic and economical decisions, and holding communication with partners. Bringing systems to work together is becoming increasingly essential for leveraging the enterprise information systems and reaching common goals. Meanwhile, semantic sharing represents the daunting barrier for making these systems work together through a more convenient global architecture and providing interoperability and reusability. This paper concentrates on studying the application of tightening together context and ontologies which can serve as formal background for reaching a suitable global enterprise environment. It brings along architecture and implementation, with special interest for the reusability of components between these systems.

1 Introduction

Enterprise Information Systems (EIS) represent a set of crucial systems used for managing enterprise and establishing its business. EIS are growing continuously with high demand of efficiency and information quality. In addition, EIS are evolving with enterprise business and challenges to bring new functionalities through wide range of accessibility portals. The perspective of improving EIS includes providing a global enterprise view over EIS with data integration initiative combined with interoperability and reusability services [Rifaieh-b 04].

Reaching a shared understanding can function as a unifying framework for the different viewpoints, and serve as the basis of communication between people, interoperability between systems, and other system engineering benefits such as reusability, reliability, and specification. In the last decade, ontologies and context played separately a major role in many AI applications, information integration for heterogeneous and distributed systems, and system engineering. On one hand, local information systems ontologies are foreseen to play a key role in partially resolving the semantic conflicts and differences that exist among systems [Bens 03]. On the other hand, the notion of context carried through views, aspects, and roles can support development process of complex systems [Arara 04].

In this paper, we are discussing the key idea of sharing semantics in the context of EIS. We strongly feel that maturity of ontology's technology, development methodologies and tools is reaching a point where it enables IT decision makers to reconsider exploitation of adapting these technologies. Likewise, the main goal of the paper includes presenting the EISCO project. We study an architecture to show where and how the pairing up between ontology and context can improve EIS. For this purpose,

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 191–198, 2005.

we present, along with the architecture, many scenarios of use and glance over the implementation issues.

2 Contextual Ontologies

The semantic sharing problem manifests, due to semantic heterogeneity, in many applications of distributed EIS. Semantic heterogeneity is classified as: heterogeneity of concepts, heterogeneity of concept structures (hierarchy), and heterogeneity of object instances. In essence, semantics sharing between users of large communities with diversified perspectives is a challenging direction of research that requires more attention. At present, concepts are expressed formally as a single representation, in the sense that the representation language is characterized by defining a unique concept and its properties as a fixed data set. In contrast to this assumption, a real world entity is unique but it can have several representations [Bens03] due to various interests, purposes, or perspectives.

Firstly, ontologies are foreseen to play a key role in partially resolving the semantic conflicts and differences that exist among representations. Secondly, ontologies themselves are not sufficient and need to be coupled with context in order to resolve the problem of multiple views and multi-representation left in semantic sharing. We strongly argue that combining the two notions of context and ontology can support the aspect of semantic sharing in both of the local view and global view [Rifaieh-a 04]. The contextual ontologies are defined at the abstraction level to take into an account the diverse points of view and multi-representation.

2.1 Contextual Ontologies Approach

Globally, the approach respects the representation for many points of view for the same concept. It provides a global view over local ontology and offers the use of many modelling representation with correspondences.

Two main assumptions are used in the approach of contextual ontologies:

- Stamping mechanism: First of all, we need to differentiate between concepts that belong to different contexts. For this reason, we recall the stamping or trade marking technique, which can be used to distinguish one representation of the same element from other representations [Arara 04].
- Relating contexts: The second issue concerning the contextual ontologies is the potential to define the semantic relationships between multi-represented concepts. In other words, we define directional bridges rules that relate the elements of a context to those of another context [Rifaieh-b 04].

2.2 Using Contextual Ontologies Within EIS

We argue, in this paper, that using the contextual ontologies as a common description for EIS conceptualization (based on multiple view of system specification) can give systems the ability to share easily more semantics. Let us consider two information systems used in an enterprise: PMIS (Project Management Information System) and HRIS (Human Resource Information System). The UML models, defined in Figure 1 and Figure 2, represent the mono-representation of each system. These systems contain concepts using with the same identifier and having the same meaning, such as *Manager* in PMIS and *Manager* in HRIS, or concepts that are different but having the same components, structure and semantically similar, such as *Engineer* in PMIS and *Developer* in HRIS. In this case, we can identify that the concepts *Manager*, *Engineer* are multi-represented in these systems.

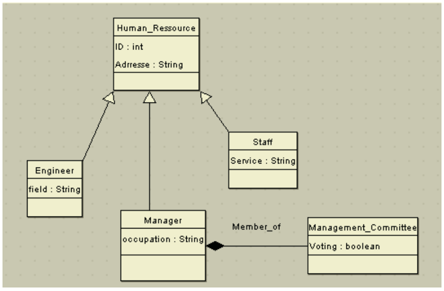


Fig. 1. A side of UML model for HRIS

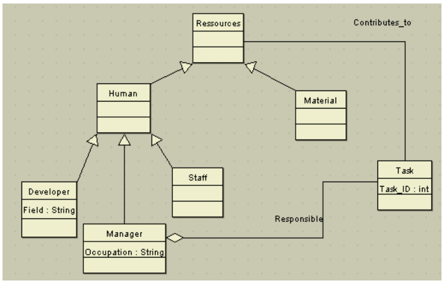


Fig. 2. A side of UML model for PMIS

We need to be able to state that two elements (e.g. concepts, roles, individuals) of two ontologies, though being contextually different, are related, because they both refer to the same object in the world. Therefore, a bridge rule asserting the identity can be expressed for these concepts. For instance, the concepts in question can help to create a new concept using some information coming from these different representations, e.g. the concept *Mangement_committee_member*. Let us refer I_1= "*John Smith*" is an instance of *Manager* in the system S_1, the ontology can offer the knowledge corresponding to I_1 in the system. With a rule R_{12} identifying that a *Manager* in S_1 is identical to a *Manager* in S_2, the system can deduce a new knowledge. Thus, one can infer that I_1 is a member of *Management_committee*.

3 EISCO Project

EISCO (Enterprise Information System Contextual Ontology) is a software engineering project, which is dedicated to exploit rich conceptual models based on contextual ontologies. Hence, these ontologies will be commonly used by several systems and applications within the enterprise. It allows sharing the same concept among several applications and representing each concept with a multiplicity of representation, such as, different roles, attributes, and instances. The notion of context is used in order to allow systems to preserve the semantics locally, whereas inter-relation between information sources is performed using coordination or bridge rules.

3.1 EISCO Architecture

As far as the impact of the EISCO server architecture will be at the enterprise level of information systems. Each running EIS should be accessible through a connector to the EISCO server in the global architecture, regardless of their horizontal, vertical or external purpose. Thus, the ESIMO server encompasses and insures a high level of services useful for the cooperation of distributed enterprise information systems. The EISCO server (Figure 3) ensures functionalities such as: (i) access to ontologies, the KB server and the inference engine; (ii) controls of the EISCO contextual ontologies and coordination rules; (iii) provides services such as resolving ambiguities (e.g. resolves ambiguity about concept, attributes, methods, etc.); (iv) provides an underlying architecture for services reusability (objects, components, patterns, etc.), interoperability, and query answering (treat and resolves global queries); (v) contains set of objects that can be remotely called by different applications; (vi) provides accessibility connectors to running applications.

3.1.1 EISCO KB Server

The EISCO knowledge base is a centralized repository for organizing system models representing ontologies (i.e. contextual ontologies) with machine-processable semantics for the dissemination of knowledge used to optimize information retrieval, interoperation and reuse. In general, a knowledge base is not a static collection of information, but a dynamic resource that may itself have the capacity to improve as a part of the inference engine results. For example, at the ontological level; an *engineer* works on a *project* having resources. At the knowledge base level, *John Smith* is an instance of *engineer* and works on the project *Security Management*.

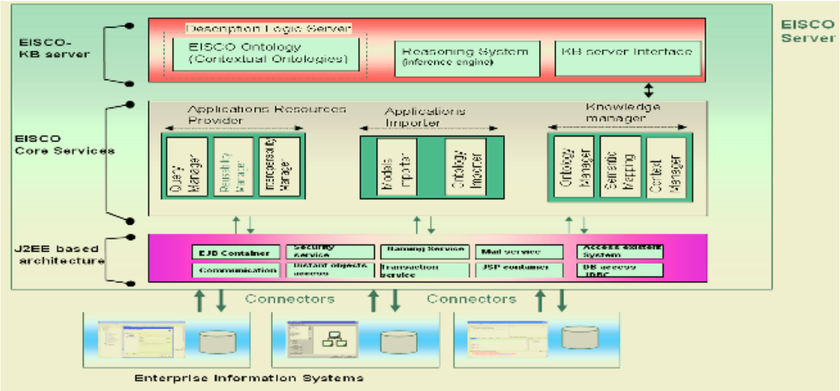


Fig. 3. EISCO Server architecture

KB Server Interface: the role of the KB Server Interface is to ensure the communication between the EISCO KB server and other components of the architecture. This interface implements all the methods for managing and maintaining the knowledge base.

Reasoning System (Inference Engine): a main advantage of using and underlying KB in our architecture is its reasoning capabilities. Therefore, the EISCO knowledge base server provides not only the access to the KB, but also the reasoning capacity over a knowledge base where results depend only on explicit semantics.

3.1.2 EISCO Core Service

The EISCO core server contains common services in the architecture platform for managing, controlling, and disseminating knowledge flow throughout the system, and for sharing services among applications (e.g., reusable objects, information, data, etc.). It includes many other functionality-oriented components such as Ontology Manager, Semantic Mapper, Context Manager, Reusability Manager, Models Importer, etc.

3.1.2.1 Applications Resources Provider

This set of components is responsible in offering connected application (i.e. systems), the basic outcome of the EISCO Server Architecture. This includes:

Query Manager: manages the life cycle of global queries and assigns to user its correspondent context query. It performs the KB Server request in order to find the similarities between the contextual concepts. According to the results, the system generates many distinct queries performed on the local systems. Finally, it ensures crossing the results to answer the global query.

Reusability Manager: provides to the developer a set of services to create a new system. The model designer attaches a set of objects and concepts to be reused afterward. The developer can express a need to a reusable object related to a concept defined in an ontology. The reusability manager then uses the set of existing bridge rules, as well as inferred links from the knowledge base, to identify similar concepts and implemented objects. Some of these objects can be reused directly, and some may be reusable with certain adaptations.

Interoperability Manager: provides interfaces in order for connected applications to cooperate and to share process and information. Therefore, making systems work together is concluded by interfacing through the Interoperability Manager, which takes into consideration how to resolve ambiguity with the help of EISCO KB.

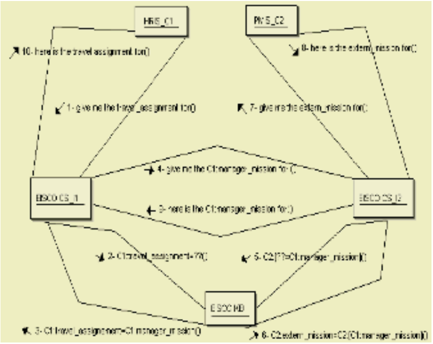

HRIS_C1
PMS_C2
10- here is the travel assignment for()
8- here is the extern_mission for()
1- give me the travel_assignment for()
7- give me the extern_mission for()
4- give me the C1:manager_mission for()
EISCO CS_1
EISCO CS_2
3- here is the C1:manager_mission for()
2- C1:travel_assignment=??()
5- C2[??=C1:manager_mission]()
EISCO KB
3- C1:travel_assignment=C1:manager_mission()
6- C2:extern_mission=C2[C1:manager_mission]()

Fig. 4. Collaboration Diagram of Interoperability Scenario

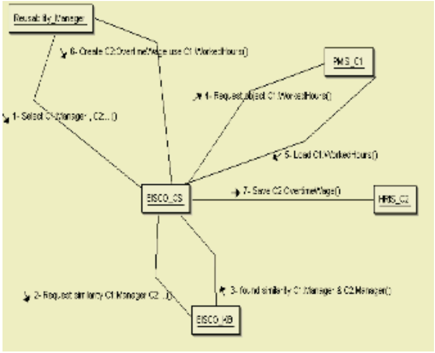

Reusability_Manager
6- Create C2:OvertimeWage use C1:WorkedHours()
PMS_C1
4- Request object C1:WorkedHours()
1- Select C1:Manager , C2...()
5- Load C1:WorkedHours()
EISCO_CS
7- Save C2:OvertimeWage()
HRIS_C2
2- Request similarity C1:Manager C2...()
3- found similarity C1:Manager & C2:Manager()
EISCO_KB

Fig. 5. Collaboration Diagram of Reusability Scenario

3.1.2.2 Applications Integrator

It ensures the input to the EISCO system. This category permits applications to be inserted in the EISCO architecture when their models or local ontologies are imported.

Models Importer: is accessible through the administrator interface to permit the import of an application or a system in the EISCO architecture. This component provides the possibility to convert the imported system model (UML, etc.) to an ontological model, accepted by the EISCO KB Server.

Ontologies Importer: consists of helping the administrator to import an ontology and to configure it to be accessible through the architecture. It helps to convert imported ontology to conform to an authorized ontology language such as OWL, DAML/OIL, etc.

3.1.2.3 Knowledge Manager

It is responsible for ensuring the management of the KB, and the use of it, by the other core services components. All of these services are accessible to the architecture administrator, and are tightly related with other components of the EISCO core services.

Ontology Manager: is used for managing ontology, helping to resolve conflicts, and keeping track of changes from previous manipulations. It manages, as well, the versioning for ontologies evolution.

Semantic Mapper: permits manipulating the semantic mapping and saving the mapping through a conventional form of ontology mapping rules. It offers the recall of these mappings on the need of EISCO CS components. It can provide, as well, a manual matching and/or semi-automatic matching to create the alignment between used ontologies.

Context Manager: considers managing the context depend information and manipulating the stamping techniques that are going to be used by the contextual ontology. It also helps the administrator of the EISCO server to assign the stamping mechanism between used ontologies.

3.1.3 EISCO Accessibility Server

EISCO Accessibility Server contains the infrastructure services providing a low-level, but robust suite of middleware services, tools, and frameworks, which simplify the development of EISCO project connections to existing EIS. More specifically, challenging software development practices such as threading, concurrency, database connectivity, object pooling, and load-balancing implementations are off-loaded from the development of EISCO, and integrated into the accessibility server. It manages the distributed computing, data flow throughout the system, data exchange, and physical accessibility (e.g., TCP/IP, CORBA, JNI). Therefore, the Accessibility Server provides, as well, connectors used to get systems plugged to the EISCO server. Each EIS should be available through a specific connector bridging it to the accessibility server.

3.2 Scenarios of Use

Two applications are presented in this section in a fairly specific and concrete form. Firstly, the semantic interoperability scenario is treated with EISCO. Secondly, a scenario about reusability with EISCO is treated to enable architect and developers to increase their efficiency and reduce the development process by using reusable objects. For these scenarios, let us use the examples of *PMIS* (Project Management Information System) Figure 1 as C1 and *HRIS* (Human Resource Information System) Figure 2 as C2.

3.2.1 Interoperability

The interoperability, by definition, is the ability for many systems to participate together for a common goal. Given two systems, interoperability can be illustrated using some methods existing in the first, for computing a value needed for the second system. We are trying to calculate in HRIS the "*manager_travel_bonus*", which occurs if a *Manager* had traveled for a number of missions. The information concerning the current missions of the *Manager* is part of PMIS, and can be computed with the "*number_extern_mission*" method. We have for certain to make the two systems collaborate to reach this goal. Therefore, the term "*travel_assignment*" is going to be translated into the term "*extern_mission*" via the contextual ontologies, which include the multi-representation aspect of the concept "*manager*". The sequence of the events is represented in Figure 4 as a collaboration diagram with the syntax of stamping mechanism of contextual ontology.

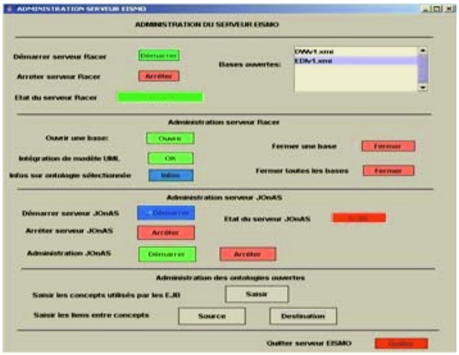
3.2.2 Reusability

Studying reusability with the EISCO Server includes finding the relation between the implemented components defined in different contexts. Let us consider that the system PMIS was implemented, it includes the model layer of the system and the implementation layer containing objects, components, etc. Let us imagine two methods: *Hours_remaining* and *Worked_hours,* developed in the context of the system PMIS (C_1). In order to help the human resource office, the enterprise decided to create a new HRIS. The development of any system starts by identifying specification and creating a conceptualization model, which should to be done with a modelling language such as UML. In order to integrate the new system in the global architecture of the enterprise, an EISCO contextual ontology should be created for the system HRIS with respect to the context (C_2). The contextual ontology should also consider the definition of the semantic similarity of the multi-represented concepts between the context C_1 and C_2. After achieving this step, we should consider how to simplify the implementation of the new system by reusing some existing components. The sequence of events is represented with the collaboration diagram in Figure 5.

3.3 Implementation of EISCO Project

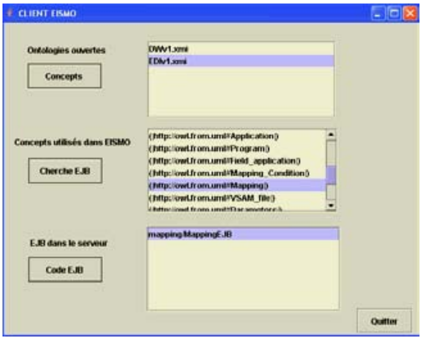
In terms of implementation, the architecture combines Knowledge Bases driven by ontological conceptualization and the J2EE platform as an implementation framework for reaching reusability and interoperability. Thus, the implementation uses Racer Description Logic reasoning system for supporting contextual ontologies. Indeed, some successful implementations of KBS, such as Racer [*Haarslev 01*] or FACT [Horrocks 00], can upload an UML model formatted in XMI and generate the ontology. It can also be used to check the coherence of a UML model [*Haarslev 01*].

The implementation uses, as well, an Open Source of the J2EE specification named JOnAS. This open application server (JOnAS) offers a container to make Enterprise Java Beans (EJB) used, within EISCO project, accessible to many EIS in the architecture. According to the scenario of reusability studied within the EISCO project, we have implemented the components of the architecture for validating the scenario. The implementation includes the development of *Administrator Graphical User Interface*, *Client Graphical User Interface, and need components of ESICO architecture* (Figure 6) and (Figure 7).



Bases ouvertes:
DWv1.xmi
EDIv1.xmi

Fig. 6. GUI for Administrator of EISCO



Ontologies ouvertes
DWv1.xmi
EDIv1.xmi

Concepts utilisés dans EISMO
(http://owl.from.uml#Application)
(http://owl.from.uml#Program)
(http://owl.from.uml#Field_application)
(http://owl.from.uml#Mapping_Condition)
(http://owl.from.uml#Mapping)
(http://owl.from.uml#VSAM_file)

EJB dans le serveur
mapping.MappingEJB

Fig. 7. GUI for Client of EISCO

4 Conclusion

In the near future, ontologies will have a direct impact on cost effectiveness and information quality. As a matter of fact, ontologies formalism and applications, as a key technology for information sharing and exchange, will certainly become great assets for enterprises. We presented in this paper a general framework that makes use of contextual ontologies. We applied, for this purpose, two scenarios of reusability and interoperability. Currently, we are finishing a case study of using the architecture in a full implemented EISs. This study covers the usefulness of contextual ontologies for data integration and data exchange platforms respectively with Data Warehouse System and Electronic Data Interchange System [Rifaieh-b 04].

References

[Bens 03] Djamal Benslimane, Ahmed Arara, "*The multi-representation ontologies: a contextual description logics approach*", In the proceeding of The 15th Conference On Advanced Information Systems Eng., Austria, 16 - 20 June, 2003, Springer-verlag.

[Arara 04] Ahmed Arara, Djamal Benslimane, *"Multiperspectives Description of Large Domain Ontologies"*, In proc. of FQAS'04.

[Rifaieh-a 04] R.Rifaieh, A.Arara, A.N.Benharkat, "*Multi-representation Ontologies in the Context of Enterprise Information Systems*", In the proceeding AMCIS'2004, New York, NY, USA, August 2004.

[Rifaieh-b 04] R.Rifaieh "*Using Contextual Ontologies for Semantic within Enterprise Information Systems*", PhD Thesis, National Institute of Applied Sciences of Lyon, Lyon, France, December 2004.

[*Haarslev 01*] *V.Haarslev, R.Möller,* "*Description of the RACER System and its Applications*", Proceedings International Workshop on Description Logics (DL-2001), Stanford, USA, 1.-3, August 2001.

[Horrocks 00] I. Horrocks, "Benchmark analysis with fact". In *Proc. of the 4th Int. Conf. on Analytic Tableaux and Related Methods (TABLEAUX 2000)*, number 1847 in Lecture Notes in Artificial Intelligence, pages 62-66. Springer-Verlag, 2000.

Mapping Fuzzy Concepts Between Fuzzy Ontologies*

Baowen Xu[1,2,3], Dazhou Kang[1], Jianjiang Lu[1,2,4], Yanhui Li[1], and Jixiang Jiang[1]

[1] Department of Computer Science and Engineering, Southeast University,
Nanjing 210096, China
[2] Jiangsu Institute of Software Quality, Nanjing 210096, China
[3] State Key Laboratory of Software, Wuhan University, Wuhan 430072, China
[4] PLA University of Science and Technology, Nanjing 210007, China
bwxu@seu.edu.cn

Abstract. Fuzzy ontology mapping is important for handling uncertain knowledge on the semantic web. However, current ontology mapping technologies are not sufficient for fuzzy ontologies. This paper proposes a framework of mapping fuzzy concepts between fuzzy ontologies. It applies the approximate concept mapping approach, extends atom fuzzy concept sets and defines the least upper bounds to reduce the searching space. It resolves the mapping problem of fuzzy concepts into finding the simplified least upper bounds for atom fuzzy concepts, and gives an algorithm for searching the simplified least upper bounds, which is fast and proved correct. The framework is efficient for mapping fuzzy concepts between fuzzy ontologies.

1 Introduction

Ontology is the basic of sharing and reusing knowledge on the semantic web. We often need represent uncertainty information [1]. The fuzzy ontologies are capable of dealing fuzzy and uncertain knowledge [2], [3]. Ontologies face the problem with respect to heterogeneity, since different systems may use different ontologies. Ontology mapping [4] is necessary to solve the problem. However, there is no published research result about fuzzy ontologies. Since fuzzy concepts often represent uncertain knowledge, approximate concept mapping is necessary. But current approximate concept mapping technologies [5], [6] are not sufficient for fuzzy ontologies.

This paper proposes a framework of mapping fuzzy concepts between fuzzy ontologies. We use the fuzzy description logics [7] to define fuzzy ontologies.

Let O be a ontology, and T be the set of atom concepts in O, then the concepts in O are defined as: $C ::= A \mid \neg C \mid C \vee D \mid C \vee D$, where $A \in T$. The interpretation of O is a pair $(\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$, where the domain $\Delta^{\mathcal{I}}$ is a set of instances, the function $\cdot^{\mathcal{I}}$ maps every fuzzy concept into $[0,1]^{\Delta^{\mathcal{I}}}$. Let C, D be fuzzy concepts, if for any $a \in \Delta^{\mathcal{I}}$, $C^{\mathcal{I}}[a] \le D^{\mathcal{I}}[a]$, then we say D subsumes C, notated as $C \sqsubseteq D$. We say D is equivalent to C, notated as $C \equiv D$, if $C \sqsubseteq D$ and $D \sqsubseteq C$. $C \sqsubset D$ means D properly subsumes C, i.e. $C \sqsubseteq D$ but C, D are not equivalent.

* This work was supported in part by the NSFC (60373066, 60425206, 90412003), National Grand Fundamental Research 973 Program of China (2002CB312000), National Research Foundation for the Doctoral Program of Higher Education of China (20020286004)

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 199–205, 2005.

2 A Framework of Mapping Fuzzy Concept

Current approximate concept mapping technologies [5, 6] only focuses on the one-to-one concept subsumption mapping between ontologies, i.e. they only consider subsumption relations between atom concepts. However, since the definition of fuzzy concept subsumption is far stricter, there are rarely one-to-one subsumption mapping between fuzzy ontologies. Therefore, it is necessary to consider subsumption relations between complex fuzzy concepts.

Let O_1 and O_2 be two fuzzy ontologies, their atom fuzzy concept sets be T_1 and T_2 which are different sets, and they are interpreted on the same domain. Our approach maps each fuzzy concept in O_1 to its approximate fuzzy concept (called approximation for short) in O_2 based on fuzzy concept subsumption. Both fuzzy concepts can be complex. There are two kinds of approximations:

Definition 1. For any fuzzy concept C in O_1, D in O_2, if $C \sqsubseteq D$, then D is a upper approximation in O_2 of C; if $D \sqsubseteq C$, then D is a lower approximation. If there are fuzzy concepts C^+ and C^- in O_2 such that for any fuzzy concept D in O_2, it is true that $C \sqsubseteq D \rightarrow C^+ \sqsubseteq D$ and $D \sqsubseteq C \rightarrow D \sqsubseteq C^-$, then C^+ and C^- are both best approximations in O_2 of C, we say C^+ is the least upper approximation, C^- is the greatest lower approximation.

The ideal objective is to find out the best approximations in O_2 for all fuzzy concepts in O_1. However, there are often massive fuzzy concepts in both O_1 and O_2. Finding and recording the best approximations in O_2 of all fuzzy concepts in O_1 offline is also impossible. We make the ideal objective a bit weaker: firstly the offline process finds and records the best approximations in O_2 of all atom fuzzy concepts in T_1; then using the recorded approximations, the online process can compute the approximations in O_2 for any fuzzy concept in O_1, when they are needed. The online process is simple: for any fuzzy concept C in O_1, firstly rewrite C into an equivalent NNF (Normal Negative Form, in which the negation operators $\neg$ only apply to atom fuzzy concepts) by equations $\neg\neg C \equiv C$, $\neg(C \wedge D) \equiv \neg C \vee \neg D$ and $\neg(C \vee D) \equiv \neg C \wedge \neg D$. Then for any atom fuzzy concept A in C:

1. If no negation operator applies to A, then replace A with A^+, otherwise replace A with A^-; the result will be an upper approximation of C;
2. If no negation operator applies to A, then replace A with A^-, otherwise replace A with A^+; the result will be a lower approximation of C.

Searching for the best approximations in O_2 for every atom fuzzy concept in T_1 is facing several problems. Two different kinds of best approximations complex the searching process. The approximations may be any fuzzy concepts in O_2 containing $\neg$, $\vee$ and $\wedge$ operators; therefore, it faces a very large searching space.

In order to avoid computing two different kinds of best approximations, we extend T_1. From Definition 1, we can easily prove that for any C in O_1, $C^- \equiv \neg(\neg C)^+$. Therefore, if we compute A^+ for any $A \in T_1$ or $\neg A \in T_1$, then all A^- can be easily computed by $A^- \equiv \neg(\neg A)^+$. We can extend T_1 so that $\forall A \in T_1 \to \neg A \in T_1$, and only searching for least upper approximation.

In order to reducing the searching space, we will eliminate $\neg$ and $\wedge$.

To eliminate $\neg$, T_2 is extended so that $\forall A \in T_2 \to \neg A \in T_2$. Since any fuzzy concept D in O_2 is a Boolean expression over T_2, and D can be turned into an equivalent NNF. We can generate any NNF from the extended T_2 by only $\vee$ and $\wedge$. $\neg$ need not be considered.

To eliminate $\wedge$, we define the least upper bounds in T_2 of A which are disjunctions of atom fuzzy concepts. Let E be a sub set of T_2, the disjunction of fuzzy concepts in E is notated as $\check{E} = \bigvee_{A_i \in E} A_i$.

Definition 2. The least upper bounds in T_2 of A is a set of disjunctions of fuzzy concepts in T_2, notated as $u(A)$, and satisfies that for $E \subseteq T_2$, $\check{E} \in u(A)$ if and only if there is no $G \subseteq T_2$ such that $A \sqsubseteq \check{G} \sqsubset \check{E}$.

If $u(A)$ is the least upper bounds in T_2 of fuzzy concept A, A^+ is the least upper approximation in O_2 of A, we can easilly get that $A^+ \equiv \bigwedge_{\check{E}_i \in u(A)} \check{E}_i$. However, the least upper bounds may still have redundancy, thus decrease the efficiency.

Definition 3. The simplified least upper bounds of A is a subset of the least upper bounds of A. $s(A)$ is the simplified least upper bounds in T_2 of A, if

$$\begin{gathered} A^+ \equiv \bigwedge_{\check{E}_i \in s(A)} (\check{E}_i); \\ \forall \check{E}_k \in s(A) \to A^+ \sqsubset \bigwedge_{\check{E}_i \in s(A)/\{\check{E}_k\}} (\check{E}_i); \\ \forall E \subseteq T_2, A \sqsubseteq \check{E} \to \exists \check{G} \in s(A), \check{G} \sqsubseteq \check{E}, |G| \leq |E|. \end{gathered} \tag{1}$$

It means the conjunction of all members in $s(A)$ is equivalent to A^+, but the conjunction of all members in any proper subsets of $s(A)$ will not be A^+, i.e. $s(A)$ uses the fewest disjunctions to compute A^+. Furthermore, $s(A)$ always choose disjunctions such that contains the fewest fuzzy concepts as its members.

The simplified least upper bounds remove the redundant members in the least upper bounds, and then simplify the expression of the least upper approximations without loosing veracity. Therefore, it only needs to find the simplified least upper bounds of a fuzzy concept to compute the least upper approximation of the fuzzy concept. It greatly reduces the searching space.

3 An Algorithm for the Simplified Least Upper Bounds

This subsection gives an algorithm for computing the simplified least upper bounds in T_2 of A. The algorithm firstly searches for the potential members of the simplified least upper bounds in a stepwise and iterative process: it seeks the independent fuzzy concepts firstly, then seeks the disjunctions of two fuzzy concepts secondly, ..., until all potential members have been found. In the end, it checks the potential members to remove redundant members. The combinatorial explosion problem may occur during the searching process, so the main objective of the algorithm is to reduce the searching space. For any vectors U, V in $[0,1]^{\Delta^{\mathcal{I}}}$, we define that

$$\begin{aligned} U \le V \ &\Leftrightarrow\ \forall a \in \Delta^{\mathcal{I}}, U[a] \le V[a]\,; \\ U < V \ &\Leftrightarrow\ (U \le V) \wedge \neg(U = V)\,; \\ \forall a \in \Delta^{\mathcal{I}}, (U \wedge V)[a] &= \min(U[a], V[a])\,. \end{aligned} \tag{2}$$

Algorithm 1. Computing the simplified least upper bounds in T_2 of A.

Input: T_1, T_2, A, $(\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$. **Output:** L. **Procedure:**

1. Let L be a ordered set for the potential members of the simplified least upper bounds, and V be a variable for a vector in $[0,1]^{\Delta^{\mathcal{I}}}$. $L = \varnothing$, $V = I(\top)$.
2. Seek the independent fuzzy concepts in T_2: for any B in T_2 such that $A \sqsubseteq B$, if $V \wedge B^{\mathcal{I}} < V$, then add B into L, and let $V = V \wedge B^{\mathcal{I}}$.
3. Generate the disjunctions of two fuzzy concepts: generate $B_1 \vee B_2$ if and only if any of the fuzzy concepts B_1, B_2 neither subsumes nor is subsumed by A, and it has neither $B_1 \sqsubseteq B_2$ nor $B_1 \sqsupseteq B_2$. Let $n = 2$.
4. Seek the disjunctions of n fuzzy concepts generated in the last step: for any disjunction $\check{E}$ such that $A \sqsubseteq \check{E}$, if $V \wedge \check{E}^{\mathcal{I}} < V$, then add $\check{E}$ into L, and let $V = V \wedge \check{E}^{\mathcal{I}}$. If $V = A^{\mathcal{I}}$, then go to step 6.
5. Generate the disjunctions of $n+1$ fuzzy concepts: generate the disjunction $\check{E}$ of $n+1$ fuzzy concepts iff for any disjunction $G \subseteq E$, if $|G| = |E| - 1$, then $\check{G}$ has been sought in the last step, and neither subsumes nor is subsumed by A. If no disjunctions can be generated, goto step 6, otherwise let $n = n+1$, go to step 4.
6. Here V must be the interpretation of the least upper approximation, and L is a superset of the simplified least upper bounds. We will validate L by Definition 3 in the following steps. Sort the members in L so that for any $\check{E}$, $\check{G}$ in L, if $|E| > |G|$, then $\check{E}$ is before $\check{G}$ in L. Actually the reverse order of the original L is suitable. Pick the first disjunction $\check{E}$ in L.
7. Compute the conjunction of members in L without $\check{E}$. If the interpretation of the result is still V, then $\check{E}$ is redundant, and we delete $\check{E}$ from L.
8. Pick the next disjunction in L, assign it to $\check{E}$ and go to step 7, until all disjunctions in L have been picked.
9. Output L, and then end the algorithm.

Theorem 1. The outputted L is the simplified least upper bounds in T_2 of A.

Proof. We use V_t, L_t to represent values of the variables V, L in the process of Algorithm 1; and V, L to represent the values of the variables V, L at the end of Algorithm 1. Let $V_k = \bigwedge_{\check{E}_i \in L, |E_i| \le k} (\check{E}_i)^{\mathcal{I}}$; c_k be the disjunctions of k fuzzy concepts which have been sought in Algorithm 1; r_k be a subset of c_k: a disjunction of fuzzy concepts is a member in r_k iff it is a member in c_k, and neither subsumes nor is subsumed by A. From Definition 1, we can get that $A^+ \equiv \bigwedge_{E_i \subseteq T_2, A \sqsubseteq \check{E}_i} \check{E}_i$. And from Algorithm, it has $V = \bigwedge_{\check{E}_i \in L} (\check{E}_i)^{\mathcal{I}}$. Since $(A^+)^{\mathcal{I}} \le V \le V_t$, from Definition 3, for any $\check{E} \in s(A)$, it must be true that $A \sqsubseteq \check{E}$, $V_t \wedge \check{E}^{\mathcal{I}} < V_t$.

Firstly, we will prove that $V = (A^+)^{\mathcal{I}}$.

1) Induction base: when $k = 1$ or 2, from step 2 and 3 of Algorithm 1,

$$V_k = \bigwedge_{\check{E}_i \in L, |E_i| \le k} (\check{E}_i)^{\mathcal{I}} = \bigwedge_{E_i \subseteq T_2, A \sqsubseteq \check{E}_i, |E_i| \le k} (\check{E}_i)^{\mathcal{I}} \tag{3}$$

For any $B \in T_2$ such that $B \in s(A)$, it has $A \sqsubseteq B$ and $V_t \wedge B^{\mathcal{I}} < V_t$, from step 2 of Algorithm 1, it is obviously that $B \in L$. For any $B \in T_2$ such that $B \notin r_1$, for any $E \subseteq T_2$ such that $\{B\} \subset G$, it is true that $\check{E} \notin s(A)$.

Similarly, we can deduce that: for any $E \subseteq T_2, |E| = 2$ such that $\check{E} \in s(A)$, it is true that $\check{E} \in L$; and for any $E \subseteq T_2, |E| = 2$ such that $\check{E} \notin r_2$, for any $G \subseteq T_2$ such that $E \subset G$, it is true that $\check{G} \notin s(A)$. So when $k = 1,2$, it is true that

$$\forall E, (E \subseteq T_2, |E| \le k, \check{E} \in s(A)) \rightarrow \check{E} \in L \tag{4}$$

$$\forall E, G, (E \subseteq T_2, |E| \le k, \check{E} \notin r_{|E|}, G \subseteq T_2, E \subset G) \rightarrow \check{G} \notin s(A) \tag{5}$$

2) Induction hypothesis: assume that when $k \le n$, Equation 3, 4 and 5 holds.

3) Induction Step: when $k = n + 1$,

For any $E \subseteq T_2, |E| \le n + 1$ such that $\check{E} \in s(A)$, assume $\check{E} \notin c_{n+1}$, then there must be a subset of E that not in r_n; from Equation 5, $\check{E} \notin s(A)$. So it must be true that $\check{E} \in c_{n+1}$, and since $A \sqsubseteq \check{E}$, $V_t \wedge \check{E}^{\mathcal{I}} < V_t$, from step 4 of Algorithm 1, it has $\check{E} \in L$. Therefore, Equation 4 holds when $k = n + 1$.

For any $E \subseteq T_2, |E| \le n + 1$ such that $\check{E} \notin r_{n+1}$, and for any $G \subseteq T_2$ such that $E \subset G$ and $|G| = n + 2$, there are only two cases:

a) If $\check{E} \notin c_{n+1}$, then there must be a subset of E not in r_n; from the induction hypothesis, $\check{G} \notin s(A)$.

b) If $\check{E} \in c_{n+1}$, then $\check{E} \notin r_{n+1}$ only if $\check{E} \sqsubseteq A$ or $A \sqsubseteq \check{E}$. $\check{E} \sqsubseteq A$ is impossible when $n \ge 1$. Because $\exists B \in E, B \sqsubseteq A$, so $B \notin r_1$ and then $\check{E} \notin c_{n+1}$. Therefore, it must be true that $A \sqsubseteq \check{E}$. Since $E \subset G$, it has $A \sqsubseteq \check{E} \sqsubseteq \check{G}$ and $|E| < |G|$, and then from Definition 3, $\check{G} \notin s(A)$.

Therefore, Equation 5 holds when $k = n + 1$.

From the definition of V_k, $V_{n+1} = V_n \wedge \bigwedge_{\check{E}_i \in L, |E_i| = n+1} (\check{E}_i)^{\mathcal{I}}$. Assume that there is $E \subseteq T_2$ such that $A \sqsubseteq \check{E}$ and $|E| = n + 1$, it is true that $V_{n+1} \wedge \check{E}^{\mathcal{I}} < V_{n+1}$ and $\check{E} \notin L$. There are also only two cases:

a) If $\check{E} \in c_{n+1}$, then since $A \sqsubseteq \check{E}$ and $V_{n+1} \wedge \check{E}^{\mathcal{I}} < V_{n+1}$, in the step 4 of Algorithm 1, $\check{E}$ must have been into L. It is a contradiction to $\check{E} \notin L$.

b) If $\check{E} \notin c_{n+1}$, then according to in the step 5 of Algorithm 1, there must be a subset of E not in r_n; from Equation 5, we can get that $\check{E} \notin s(A)$. Since $A \sqsubseteq \check{E}$, from Equation 1, it should exists $\check{G} \in s(A)$ such that $\check{G} \sqsubseteq \check{E}, |G| \le |E|$. From Equation 4, it has $\check{G} \in L$, so $V_{n+1} \le \check{G}^{\mathcal{I}}$, and then $V_{n+1} \wedge \check{E}^{\mathcal{I}} < V_{n+1}$ is impossible.

Therefore, for any $E \subseteq T_2$ such that $A \sqsubseteq \check{E}$ and $|E| = n+1$, it is true that $V_{n+1} \wedge \check{E}^{\mathcal{I}} < V_{n+1} \to \check{E} \in L$, and then $\check{E} \notin L \to V_{n+1} \le \check{E}^{\mathcal{I}}$. So it must be true that $V_{n+1} = V_n \wedge \bigwedge_{\check{E}_i \in L, |E_i| = n+1} (\check{E}_i)^{\mathcal{I}} = V_n \wedge \bigwedge_{E_i \subseteq T_2, A \sqsubseteq \check{E}_i, |E_i| = n+1} (\check{E}_i)^{\mathcal{I}}$.

Substitute V_n into it, $V_{n+1} = \bigwedge_{\check{E}_i \in L, |E_i| \le n+1} (\check{E}_i)^{\mathcal{I}} = \bigwedge_{E_i \subseteq T_2, A \sqsubseteq \check{E}_i, |E_i| \le n+1} (\check{E}_i)^{\mathcal{I}}$

Therefore, Equation 3 holds when $k = n+1$.

From 1), 2), 3), for any $k \ge 1$, Equation 3, 4 and 5 holds.

Since $V = V_{|T_2|}$, $\check{E}_i \in L \to E_i \subseteq T_2, A \sqsubseteq \check{E}_i$ and $E_i \subseteq T_2 \to |E_i| \le |T_2|$, so from Equation 3, $V = \bigwedge_{\check{E}_i \in L} (\check{E}_i)^{\mathcal{I}} = \bigwedge_{E_i \subseteq T_2, A \sqsubseteq \check{E}_i} (\check{E}_i)^{\mathcal{I}} = (A^+)^{\mathcal{I}}$.

Then we can go on proving $L = s(A)$. From $V = (A^+)^{\mathcal{I}}$, we can get that $\bigwedge_{\check{E}_i \in L} \check{E}_i \equiv A^+$. From step 7 and 8 of Algorithm 1, we can get that $\forall \check{E}_k \in L \to A^+ \sqsubset \bigwedge_{\check{E}_i \in L/\{\check{E}_k\}} (\check{E}_i)$. For any $E \subseteq T_2$ such that $A \sqsubseteq \check{E}$, from Equation 1, it should exists $\check{G} \in s(A)$ such that $\check{G} \sqsubseteq \check{E}, |G| \le |E|$. Then from Equation 4, it has $\check{G} \in L$. So we can get that $\forall E \subseteq T_2, A \sqsubseteq \check{E} \to \exists \check{G} \in L, \check{G} \sqsubseteq \check{E}, |G| \le |E|$.

From Definition 3 and the above equations, we can deduce that $L = s(A)$, i.e. L is the simplified least upper bounds in T_2 of A. □

Algorithm 1 is fast and efficient in most cases. Only disjunctions such that may be in $s(A)$ are generated in step 5 of each loop, and if a disjunction is generated, then no its supersets will not be considered. This greatly reduces the searching space. Furthermore, only disjunctions such that actually lower the value of V are added into L. Validating members in L is fast too.

When mapping two fuzzy ontologies O_1 and O_2 in practice, we first find the set of common instances of them and use them to build a $(\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$. For any atom fuzzy concept A in O_1 or O_2, we use the algorithm to find its least upper approximations. Then use $A^- \equiv \neg(\neg A)^+$ to get the greatest lower approximations. That is the offline process. After that, the online process proposed in Section 3.1 is ready to map fuzzy concepts between O_1 and O_2 according to requirements from users.

4 Conclusions

This paper proposes a framework of mapping fuzzy concept between fuzzy ontologies. It applies the approximate concept mapping approach, extends atom fuzzy concept sets and defines the least upper bounds to reduce the searching space of fuzzy concept mapping. The framework resolves the mapping problem of fuzzy concept into finding the simplified least upper bounds for atom fuzzy concept, and gives an algorithm for searching the simplified least upper bounds of fuzzy concept. The algorithm is fast and correct. The framework is efficient for mapping fuzzy concepts between fuzzy ontologies.

References

1. Widyantoro, D.H., Yen, J.: A Fuzzy Ontology-based Abstract Search Engine and Its User Studies. In: Proceedings of the 10th IEEE International Conference on Fuzzy Systems, Melbourne, Australia (2001) 1291–1294
2. Quan, T.T., Hui, S.C., Fong, M., Cao, T.H.: Automatic Generation of Ontology for Scholarly Semantic Web. In: Proceedings of the International Semantic Web Conference, Hiroshima, Japan (2004) 726–740
3. Parry, D.: A fuzzy ontology for medical document retrieval. In: Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation, Dunedin, New Zealand (2004) 121–126
4. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. The Knowledge Engineering Review, vol. 18, no.1 (2003) 1–31
5. Stuckenschmidt, H.: Approximate information filtering on the semantic web. In Proceedings of the 25th German Conference on Artificial Intelligence, Lecture Notes in Artificial Intelligence, Springer (2002) 114–128
6. Akahani, J., Hiramatsu, K., Satoh, T.: Approximate Query Reformulation based on Hierarchical Ontology Mapping. International Workshop on Semantic Web Foundations and Application Technologies (2003) 43–46
7. Straccia, U.: Reasoning within fuzzy description logics. Journal of Artificial Intelligence Research, no. 14 (2001) 137–166

Similarity Estimation of 3D Shapes Using Modal Strain Energy

Soo-Mi Choi and Yong-Guk Kim

School of Computer Engineering, Sejong University, Seoul, Korea
{smchoi,ykim}@sejong.ac.kr

Abstract. Shape comparison between 3D models is essential for shape recognition, retrieval, classification, etc. In this paper, we propose a method for comparing 3D shapes, which is invariant under translation, rotation and scaling of models and is robust to non-uniformly distributed and incomplete data sets. A modal model is constructed from input data using vibration modes and then shape similarity is evaluated with modal strain energy. The proposed method provides global-to-local ordering of shape deformation using vibration modes ordered by frequency. Thus, we evaluated similarity in terms of global properties of shape without being affected localized shape features using ordered shape representation and modal strain energy.

1 Introduction

Comparing 3D models is an essential task for an image system in which it has to index, search and classify diverse shapes. Previous studies on shape comparison have often focused on 2D images, where curvature, contour curve, silhouette, color and texture of the shape, or wavelet transformation or Fourier transformation, have been typically used as the mean of comparing the shapes. Recent efforts of 3D model shape comparison are typically to extend previous 2D image models into 3D models. However, since such cases have some limitation because of dimension transformation, it is necessary to have new way of comparing the 3D shapes. The early 3D models are: Extended Gaussian Images and Harmonic Shape Images. The other models are Generalized Cylinders, Shock Graphs, Medial Axes, and Skeletons, in which structured data descriptor is adopted for comparing the shape. However, those models cannot be used for comparing a certain models that contain holes and moreover computation complexity is also high since 2D skeleton is extended into 3D case. Recently, some 3D models utilize their geometrical and topological characteristics in measuring similarity or matching between two shapes [1,2].

Mass-spring model and Finite-element method are based upon physical characteristics. For instance, Christensen [3] shows that it is possible to deform the shape using free form deformation model by setting up a mass-spring grid consisting of 8 nodes, and Nedal [4] proposes a model by which deformation process of human muscles can be simulated in real time. Gourret [5] uses finite element method in modeling the interaction between human hand and a deformable object. Chen and Zeltzer [6] construct 20 nodes consisting of brick elements and use parabolic interpolation functions for their fine element modeling of muscle deformation. Essa et al. [7] employ dynamic finite element method in analyzing moving trajectory. They use superquadric

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 206–212, 2005.

finite elements consisting of 27 nodes and dynamic equilibrium equation. Bro-Nielson et al. [8] propose that the finite element method using linear interpolation and four surface elements can be used in modeling human skin deformation for artificial surgery. Zhu et al. [9] introduce a bio-mechanical model which is based upon voxel representation for muscle deformation using finite element method and volume rendering. Our brief survey suggests that the 3D shape comparison methods based upon the geometrical shape or object characteristics provide an important standard by which other models can be compared to it. This paper presents a new 3D comparison method in which modal strain energy required during deformation process between modal models is used to measure 3D shapes.

The rest of this paper is organized as follows. Section 2 describes modal analysis briefly as a mathematical foundation and Section 3 describes shape similarity estimation using modal strain energy. Experimental results and discussion are given in Section 4 and some conclusions and future works are given in Section 5.

2 Modal Analysis

The vibration mode shapes and frequencies are properties of any vibrating system and can be determined analytically using modal analysis. The mode shape vectors are derived from the equilibrium equation for simulating the dynamic behavior of an object, and here they are the generalized eigenvectors of the dynamic equilibrium equation without damping:

$$MU'' + KU = 0 \tag{1}$$

where U is a $3n \times 1$ vector of the (Δx , Δy , Δz) displacements of the n nodal points relative to the object's center of mass. M and K are $3n \times 3n$ matrices describing the mass and material stiffness, respectively. Eq. (1) may be interpreted as assigning a certain mass to each nodal point and a certain material stiffness between nodal points without damping. From Eq. (1), the generalized eigenproblem is obtained and yields n eigensolutions: $(\omega_1^2, \phi_1), (\omega_2^2, \phi_2), \cdots, (\omega_n^2, \phi_n)$.

$$K\phi = \omega^2 M\phi \tag{2}$$

The vector ϕ_i is called the i^{th}-mode shape vector, and ω_i is the corresponding frequency of vibration. The columns of the modal matrix Φ are the generalized eigenvector ϕ_i of M and K and the eigenvalue ω_i^2 is stored on the diagonal of matrix Ω^2 in increasing eigenvalue order $(\omega_1^2 \le \omega_2^2 \le \cdots \le \omega_n^2)$. All the eigenvectors are M -orthonormalized. Hence

$$\phi_i^T M\phi_j = \begin{cases} 1 & i = j \\ 0 & i \ne j \end{cases} \tag{3}$$

$$\Phi^T M\Phi = I, \quad \Phi^T K\Phi = \Omega^2 \tag{4}$$

Eq. (1) can be transformed into a form that is not only less costly but also allows a closed-form solution by mode superposition [10]. To diagonalize the equation, the modal matrix Φ is used. U is transformed into modal displacements $\widetilde{U}$ by

$U = \Phi\widetilde{U}$. Because M and K are normally symmetric positive definite, they are also diagonalized by Φ . As a result Eq. (1) can be rewritten as:

$$\begin{gathered}\widetilde{M}\widetilde{U}'' + \widetilde{K}\widetilde{U} = \widetilde{F} \\ \widetilde{M} = \Phi^T M\Phi,\ \widetilde{K} = \Phi^T K\Phi,\ \widetilde{F} = \Phi^T F\end{gathered} \tag{5}$$

The mode shape vectors form an orthogonal object-centered coordinate system for describing feature locations. That is, each feature point location is uniquely described in terms of how it projects onto each eigenvector.

3 Shape Similarity Estimation Using Modal Strain Energy

In our modeling framework, an object's shape is described by modal deformations within each free vibration mode. For a given set of 3D points, first we create an initial superellipsoid centered at the center of mass and rotate it into the principal axes, and then the superellipsoid is triangulated. Finite element nodes are superimposed on the initial mesh to achieve a relationship between the element displacements at any point and the element nodal point displacements directly. Where n is the number of finite element nodes and q_{ik} is the interpolation coefficients that satisfy when the interpolation function h_i 's value is one at corresponding node i and zero at all other nodes.

$$H = \begin{bmatrix} h_1 \cdots h_n & 0 \cdots 0 & 0 \cdots 0 \\ 0 \cdots 0 & h_1 \cdots h_n & 0 \cdots 0 \\ 0 \cdots 0 & 0 \cdots 0 & h_1 \cdots h_n \end{bmatrix} \tag{6}$$

where $h_i(x,y,z) = \sum_{k=1}^{n} q_{ik}\, g_k(x,y,z)$ and $g_i(x,y,z) = e^{-\left[(x-x_i)^2/2\sigma_x^2 + (y-y_i)^2/2\sigma_y^2 + (z-z_i)^2/2\sigma_z^2\right]}$

The 3D mass matrix M can be computed directly from the interpolation matrix H and ρ is the mass density.

$$M = \int_V \rho H^T H \, dV \tag{7}$$

The 3D stiffness matrix K is calculated by Eq. (8) where B is the strain displacement matrix and C is the material matrix.

$$K = \int_V B^T C B \, dV \tag{8}$$

The strain displacement matrix B is obtained by appropriately differentiating and combining rows of the element interpolation matrix H . In the Eq. (8), the material matrix C expresses the material's particular stress-strain law. We here employ the generalized stress-strain matrix for 3D isotropic materials [10]. When considering the free undamped behavior of the object, the motion equation is decoupled by solving the generalized eigenvalue problem ($K\phi = \omega^2 M\phi$) for the mass and stiffness matrices. The resulting eigenvectors describe the vibration mode shapes of the model and the eigenvalues are their corresponding frequencies of vibration. These eigenvectors are stored in the Φ matrix, and comprise the modal transformation. The non-rigid motion is described by modal displacements $\widetilde{U}$.

Modes provide global-to-local ordering of shape deformation that allows us to select which types of deformations are to be used to reconstruct the object. In general, low-frequency modes describe global deformation, while higher-frequency modes describe more localized shape deformation. By discarding high-frequency modes, the required computational time can be reduced without loss of accuracy. Moreover, the reconstructed shapes are relatively robust for input data because of the stability of low-frequency modes.

Once the modal models are reconstructed, we can compute the modal strain energy incurred by deformations. The strain energy associated with the i$^{\text{th}}$ mode is computed by Eq. (9). This can then be used as a similarity metric between 3D shapes. Fig. 1 shows the overall procedure for similarity estimation of 3D shapes using the modal strain energy.

$$E_{\text{mode}\,i} = \frac{1}{2}\tilde{u}_i^2 \omega_i^2 \tag{9}$$

The modal strain does not satisfy one of the three axioms for a similarity metric (i.e. minimality, symmetry, and triangle inequality). While it satisfies minimality and tiangle inequality, the strain energy does not satisfy symmetry. The strain energy is not symmetric for shapes of differing sizes. If the scales of two objects A and B differ, then the strain energy needed to align A with B may differ from that needed to align B with A.

$$\begin{aligned} \text{Minimality: } & \delta(A,B) \ge \delta(A,A) = 0 \\ \text{Symmetry: } & \partial(A,B) = \delta(B,A) \\ \text{Triangle inequality: } & \delta(A,B) + \delta(B,C) \ge \delta(A,C) \end{aligned} \tag{10}$$

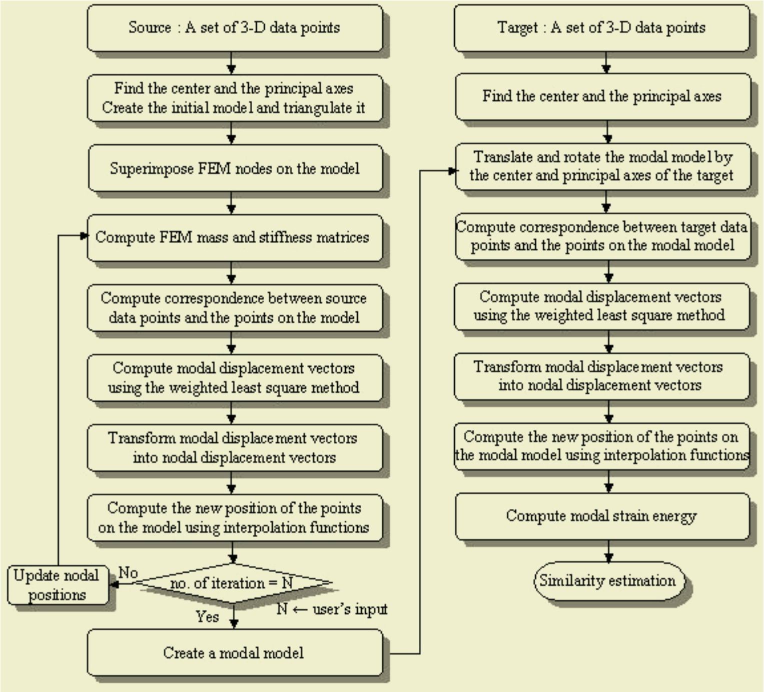

Source : A set of 3-D data points
Find the center and the principal axes
Create the initial model and triangulate it
Superimpose FEM nodes on the model
Compute FEM mass and stiffness matrices
Compute correspondence between source data points and the points on the model
Compute modal displacement vectors using the weighted least square method
Transform modal displacement vectors into nodal displacement vectors
Compute the new position of the points on the model using interpolation functions
no. of iteration = N
No
Update nodal positions
Yes
N ← user's input
Create a modal model
Target : A set of 3-D data points
Find the center and the principal axes
Translate and rotate the modal model by the center and principal axes of the target
Compute correspondence between target data points and the points on the modal model
Compute modal displacement vectors using the weighted least square method
Transform modal displacement vectors into nodal displacement vectors
Compute the new position of the points on the modal model using interpolation functions
Compute modal strain energy
Similarity estimation

Fig. 1. Overall procedure for similarity estimation of 3D shapes using modal strain energy

4 Results and Discussion

Our similarity estimation between 3D shapes was applied for data sets from meshes and image contours. In the mesh case, points were sampled from a number of existing surfaces. Fig. 2 (left) shows some examples. The recovered modes and computing times are shown in Table 1 for each of the examples. This experiment shows that the reconstructed modal model is relatively robust for partially missing data. Such robustness comes from the stability of low-order mode shapes and the unification of geometric and physical models. By describing point locations using free vibration modes, it is also easy to measure the similarity between different objects. Fig. 3 depicts the use of modal strain energy for comparing a sample object with nine other objects. Fig. 3(a) shows some reconstructed modal models for the shape comparison. The modal strain energy that results from deforming the sample to the shape of each object is illustrated in the graphs Fig. 3(b) and (c).

Table 1. Recovered modes and computing times required to reconstruct object's surfaces

Input models	Modal models		
No. of vertices	No. of vertices	No. of modes	Processing time(sec)
(a) 384	762	258	130
(b) 576	762	258	145
(c) 480	762	258	134
(d) 210	762	258	123

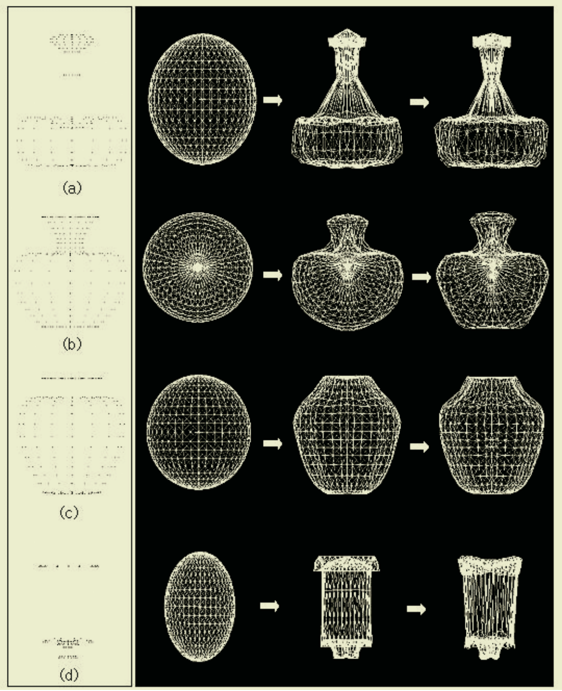
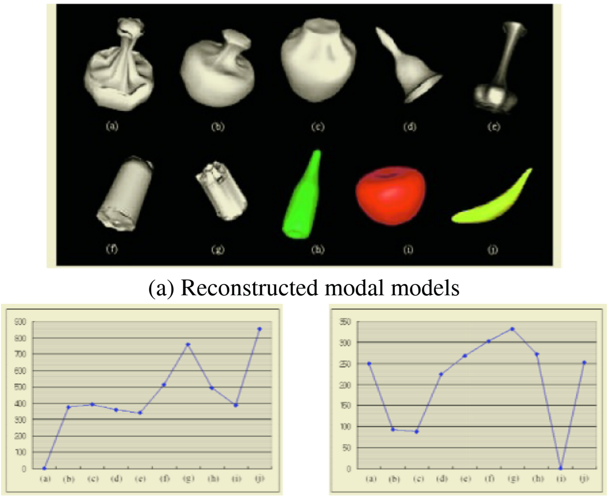


Fig. 2. Reconstruction of modal models from input data sets



(a) Reconstructed modal models

(b) Strain energy graph: sample (a) (c) Strain energy graph: sample (i)

Fig. 3. Similarity estimation between 3D shapes using the modal strain energy

5 Conclusions and Future Works

This paper presents a new method for 3D shape comparison using the modal strain energy. The present method is invariant under translation, rotation and scaling of models and is robust to non-uniformly distributed and incomplete data sets. A modal model is constructed from input data using vibration modes and then shape similarity is evaluated with modal strain energy. The proposed method provides global-to-local ordering of shape deformation using vibration modes ordered by frequency. Thus, we evaluated similarity in terms of global properties of shape without being affected localized shape features using the ordered shape representation and the modal strain energy. Our experimental results suggest that the model is robust against global and local shape deformation by comparing the other similar models.

Since the present model represents all 3D shapes as the closed forms, it can not deal with, for instance, a sphere-shape topology. The reason of adopting such scheme is that it is able to model the case that has incomplete data set. And, yet, the model has to be improved, by which one can choose and vary the topology of the object.

Acknowledgements

This work was supported by the Korea Research Foundation Grant. (KRF-2002-003-D00308)

References

1. Hilaga, M., Shinagawa, Y., Kohmura, T., Kunii, T.L.: Topology Matching for Fully Automatic Similarity Estimation of 3D Shapes, The proceeding of SIGGRAPH (2001) 203-212
2. Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D.: Matching 3D Models with Shape Distribution, Proc. Shape Modeling Int'l (2001)

3. Christensen, J., Marks, J., Ngo, J.T.: Automatic motion synthesis for 3D mass-spring models, The Visual Computer, Vol. 13 (1997) 20-28
4. Nedel, L.P., Thalmann, D.: Real time muscle deformation using mass-spring systems, Proceedings of Computer Graphics International (1998) 156-165
5. Gourret, J.P., Thalmann, N.M., Thalman, D.: Simulation of object and human skin deformations in a grasping task", Computer Graphics Proceedings, ACM SIGGRAPH (1989) 21-30
6. Chen, D., Zeltzer: Pump it up: computer animation of a biomechanically based model of muscle using the finite element method, Computer Graphics Proceedings, ACM SIGGRAPH (1992) 89-98
7. Essa, I., Sclaroff, S., Pentland, A.: A unified approach for physical and geometric modeling for graphics and animation, Proceedings of Eurographics, Vol. 11 (1992) 129-138
8. Bro-Nielsen, M., Cotin, S.: Real-time volumetric deformable models for surgery simulation using finite elements and condensation, Proceedings of Eurographics, Vol. 15 (1996) 57-66
9. Zhu, Q.H., Chen, Y., Kaufman, A.B.: Real-time Biomechanically-based Muscle Volume Deformation using FEM, Computer Graphics Forum, Vol. 17, No. 3 (1998) 275-284
10. Bathe, K.: Finite Element Procedures in Engineering Analysis, Prentice-Hall (1982)

3D-Based Synthesis and 3D Reconstruction from Uncalibrated Images

Sang-Hoon Kim[1], Tae-Eun Kim[2], Mal-Rey Lee[3], and Jong-Soo Choi[1]

[1] Department of Image Engineering Graduate School of Advanced Imaging Science,
Multimedia, and Film Chung-Ang University, 221 Huksuk-Dong, DongJak-Ku, Seoul, Korea
{sh_kimsh,jschoi}@imagelab.cau.ac.kr
[2] Department of Multimedia, Nam-Seoul University,
21 Maeju-ri, Seonghwan-eup, Chonan,Korea
tekim@nsu.ac.kr
[3] Department of Electronics and Information Engineering,
Chon-Buk National University, Jeonju,Chonan,Korea
mrlee@mail.chonbuk.ac.kr

Abstract. In this paper, we propose a new camera calibration method for the 3D-based image synthesis and 3D reconstruction. We improve the problem as changing the principle point for obtaining the linear equation. According to the error rate, we adapt the non-linear method that minimizes the intrinsic parameters. Namely, it minimizes the intrinsic parameters error with maintaining the computational conciseness. As a result, we can find optimized camera intrinsic parameters and adapt to image synthesis and reconstruction. Experimental results show the performance of the proposed method is the better than the previous. We also demonstrate examples of the 3D-based image synthesis and 3D reconstruction from uncalibrated images.

1 Introduction

One of the main research field in computer vision is the matching of stereoscopic images. This matching enables the building up of a 3D surface of the scene. Camera calibration is the fundamental task for the 3D-based synthesis and 3D reconstruction from the correspondences. It is divided into two methods roughly. One is the auto calibration method and the other is using 3D data such as the pattern information. The former is more progressive method, because there is no constraint or previous information of the scenes. Therefore, auto-calibration method has a difficult progress dealing with the complicated non-linear equations. By reason of this, it gives some restricted condition on the intrinsic camera parameters in recent researches. In this case, we can solve the linear equations instead of complicated non-linear equations. Auto-calibration is the computation of camera internal calibration and/or metric properties of the scene from a set of uncalibrated images. [1][2][3][4] The original auto-calibration method based on kruppa's equations was restricted to cameras with fixed internal parameters, and early work in this area maintained this restriction. However, this constraint occur the error to the camera intrinsic parameters. We present a simple approach to auto-calibration for the purpose of reducing the error. We formulate a constraint in terms of the variable principle points and improve the problem as changing the principle point for obtaining the linear equation. According to the error rate, we adapt the non-linear method that minimizes the intrinsic parameters. Namely, it

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 213–218, 2005.

minimizes the intrinsic parameters error with maintaining the computational conciseness. Experimental results show the performance of the proposed method is better than the previous. We also demonstrate examples of the 3D-based image synthesis and the 3D reconstruction.

2 Matching Algorithm

We have modified the previous corner matching algorithm to apply in image sequences more efficiently. [5][7] The correlation and strength of match measure equations presented in the paper by Zhang et. al. and calculate some correspondence between corners in the two images.[5][6] A correlation window of size (2n+1)× (2m+1) is centered at each corner detected in the first of two images. A rectangular search area of size 2du+1) × (2dv+1) is placed around this point in the second image and for all the corners falling inside this area. we implemented n=7, m=7 for the correlation window, du=(width)/4, dv=(height)/4 for the search window and the threshold was chosen to be 0.8. The correlation based matching correspondences method has a problem of false matches. Therefore more important thing in the correspondence problem is to remove the false matches.

$$Score(m_1,m_2)=\frac{\sum_{i=-n}^{n}\sum_{j=-m}^{m}\left[I_1(u_1+i,v_1+j)-\overline{I_1(u_1,v_1)}\right]\times\left[I_2(u_2+i,v_2+j)-\overline{I_2(u_2,v_2)}\right]}{(2n+1)(2m+1)\sqrt{\sigma^2(I_1)\times\sigma^2(I_2)}} \tag{1}$$

$$\overline{I_k(u,v)}=\sum_{i=-n}^{n}\sum_{j=-m}^{m}I_k(u+i,v+j)/\left[(2n+1)(2m+1)\right] \tag{2}$$

$$\sigma(I_k)=\sqrt{\frac{\sum_{i=-n}^{n}\sum_{j=-m}^{m}I_k^2(u,v)}{(2n+1)(2m+1)}-\overline{I_k(u,v)}} \tag{3}$$

Our improved method for the correspondence extracts the motion vectors in the four divided areas. From the extracted motion vectors, we remove the false matches. A representative motion vector in each area can be compared with the motion vectors. If the difference of the motion vector length is over ±30%, and direction is over ±5°, we decide the correspondences are false matches and remove them. It is more efficient to find correspondences in sequences with small motion.

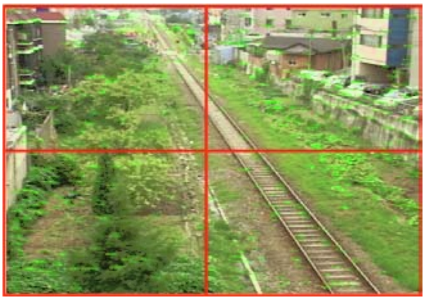


Fig. 1. Image division for the representative motion vectors

3 Proposed Calibration Approach

General camera calibration methods for inducing the linear equations can generate the errors instead of reducing the computation costs by enforcing the strong constraint. [1][2][3][4][8] We improve the problem as changing the principle point for obtaining the linear equation. According to the error rate, we adapt the non-linear method that minimizes the intrinsic parameters error. Namely, it minimizes the intrinsic parameters error with maintaining the computational conciseness. Generally, it is assumed that the skew and principle points are zero.[10][11] This disregards an error by the camera distortion. Our proposed method is represented that minimizes the error differs in each camera. Camera parameters can be recovered from uncalibrated images in the following stages.

1. Set the matrix C for varying the principle points

$$P' = CP, \quad C = \begin{bmatrix} 1 & 0 & -o_{xi} \\ 0 & 1 & -o_{yi} \\ 0 & 0 & 1 \end{bmatrix}, \quad \begin{matrix} -d \le o_{xi} \le d \\ -d \le o_{yi} \le d \end{matrix} \tag{4}$$

2. Compute the absolute dual quadric □ [8][10]

$$\begin{aligned} &(P'^{(1)} Q^*_\infty P'^{(2)^T}) = 0 \\ &(P'^{(1)} Q^*_\infty P'^{(3)^T}) = 0 \\ &(P'^{(2)} Q^*_\infty P'^{(3)^T}) = 0 \\ &(P'^{(1)} Q^*_\infty P'^{(1)^T}) = (P'^{(2)} Q^*_\infty P'^{(2)^T}) \end{aligned} \tag{5}$$

3. Decision the error rate from the Error Function E

$$\mathbf{E(o_x, o_y)} = \sum_{i=1}^{n} \left\| \mathbf{P_i Q^*_\infty P_i^T} - D(\mathbf{P_i Q^*_\infty P_i^T}) \right\|_F \tag{6}$$

4 Experimental Results

Experimental results show the good matching result with a subpixel error. It is more efficient matching method in images sequence with a small motion. It has found the more correspondence with a small error. Figure.5 shows the feature points found 1803 points in figure 5(a), 1951 points in figure (b), 2019 points in figure (c), 2163 points in figure (d).

Table 1. Results of the figure 2

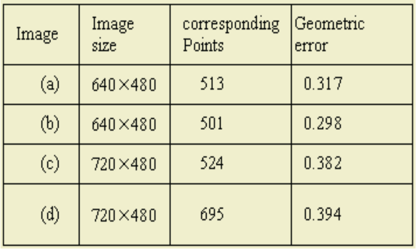
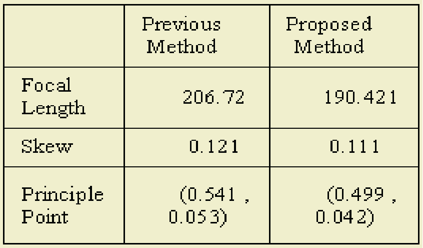


Image	Image size	corresponding Points	Geometric error
(a)	640×480	513	0.317
(b)	640×480	501	0.298
(c)	720×480	524	0.382
(d)	720×480	695	0.394

Table 2. Comparing with the previous method



	Previous Method	Proposed Method
Focal Length	206.72	190.421
Skew	0.121	0.111
Principle Point	(0.541 , 0.053)	(0.499 , 0.042)

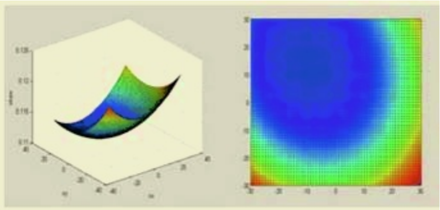


Fig. 2. Skew values as changing (ox_i, oy_i)

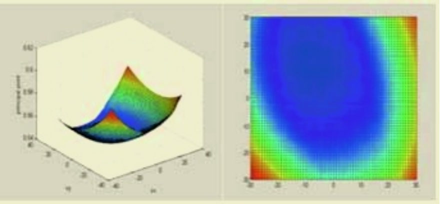


Fig. 3. Principle Points as changing (ox_i, oy_i)

We measured the intrinsic parameters error as changing the o_{xi}, o_{yi} from -30 to 30 by the propose method. As shown in figure 5 and 6, the intrinsic parameters converge on the some area. It found the intrinsic parameters at the principle point (-7,13) with a minimum error computed from the proposed error function. The table 2. represents camera parameters found in images and shows more accuracy than the previous method.

$$err_{reprojection} = \sqrt{(x_i - x_i^{'})^2 + (y_i - y_i^{'})^2}$$

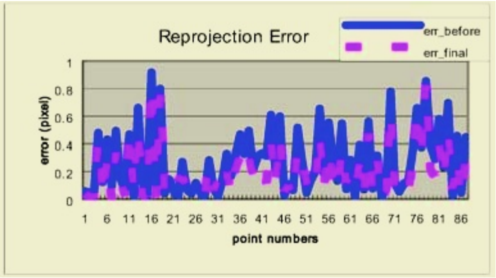
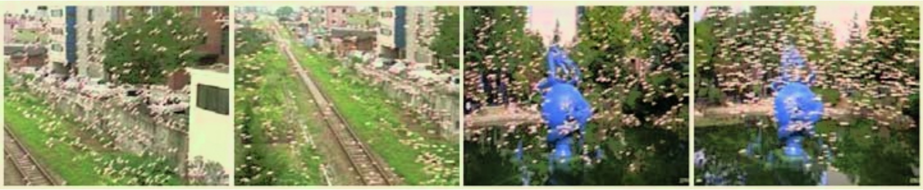


Fig. 4. Reprojection error

Table 3. Comparing with the previous method

Avg. of Reprojection error	Previous method	Proposed method
(pixel)	0.2677	0.1993

Figure.4 shows the reprojection error comparing with the previous method. Table 3. shows the average of reprojection error.



(a) Rail load Images (b) Status Images

Fig. 5. Correspondences between images

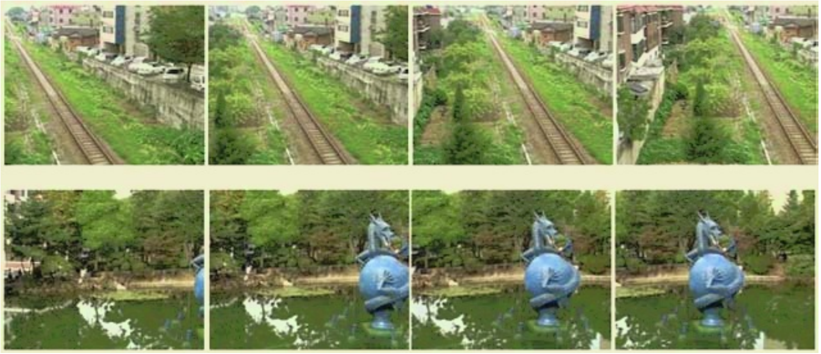


Fig. 6. Input Images

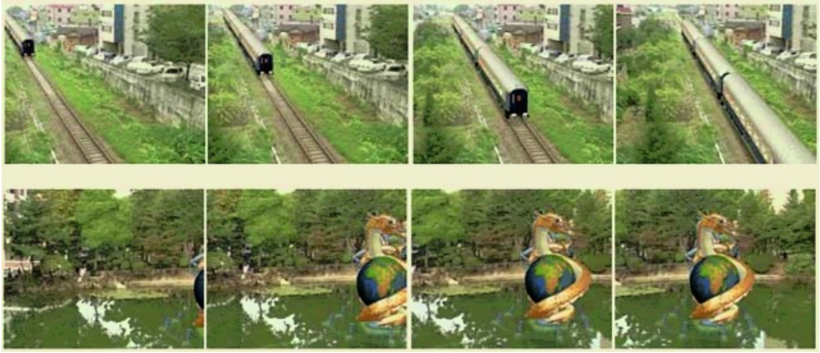


Fig. 7. Synthesis of the 3D model

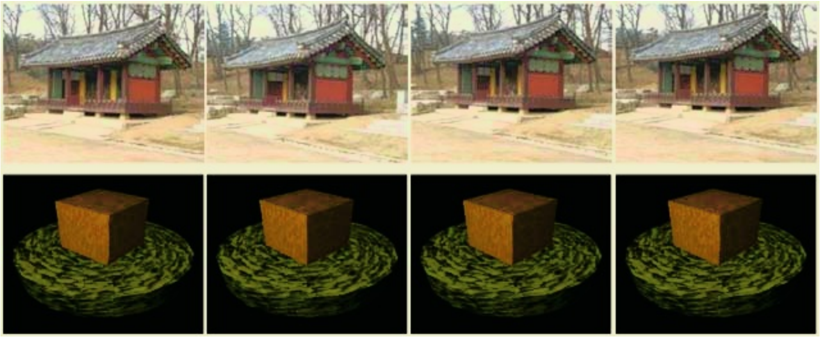


Fig. 8. Input Images

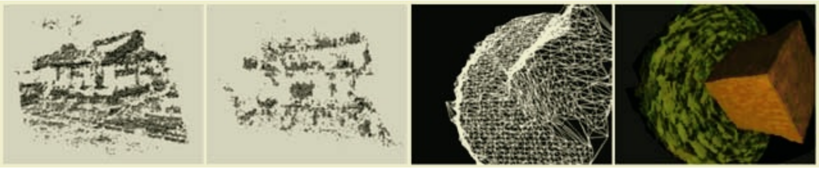


Fig. 9. 3D Reconstruction

5 Conclusion

In this paper, we have presented a camera calibration method of image sequences based on the projective factorization which is improved by the error minimization method.[11] It is possible that the principle points are changed during the calibration process. Basically, it need not to induce the complicated equations and minimizes the intrinsic parameters error with maintaining the computational conciseness. Our experiments show that the proposed approach allows to obtain the robust estimates of camera intrinsic parameters in a computational simple way that can be easily implemented in practice. We also demonstrate examples of the 3D-based image synthesis and 3D reconstruction from uncalibrated images.[13]

References

1. B. Triggs, "Autocalibration and the absolute quadric", *Proc. Conference on Computer Vision and Pattern Recognition, IEEE Computer Soc.* Press, pp. 609-614, 1997
2. A. Heyden, K. Astrom, "Euclidean Reconstruction from Constant Intrinsic Parameters" *Proc. 13th International Conference on Pattern Recognition, IEEE Computer Soc*. Press, pp. 339-343, 1996
3. R. Hartley, J.L. Mundy, A. Zisserman, and D. Forsyth (eds.), "Euclidean reconstruction from uncalibrated views", *Applications of Invariance in Computer Vision, Lecture Notes in Computer Science*, Vol. 825, Springer-Verlag, pp. 237-256, 1994
4. Marc Pollefeys, "Tutorial on 3D modeling from images", *Dublin, Ireland In conjunction with ECCV, Lecture Notes*, CH. 6, 26 June 2000
5. Zhengyou ZHANG, "A Robust Technique for Matching Two Uncalibrated Images +Through the Recovery of theUnknown Epipolar Geometry", *Technical Report*, May, 1994
6. P. Smith, D. Sinclair, R. Cipolla and K. Wood, "Effective corner matching", *BMVC*, September 1998
7. C. Harris and M. Stephens, "A combined corner and edge detector", *Fourth Alvey Vision Conference*, pp.147-151, 1988
8. R. Hartley and A. Zisserman "Multiple View Geometry in computer vision" *Cambridge Univ. Press*, CH. 5,6,8,11,18, 2000
9. Longuete Higgins "A computer algorithm for reconstructing a scene from two projections", *Natur*e, vol. 293, pp. 133-135, September 1981
10. R. Deriche, Z. Zhang, Q.T. Luong and O. Faugeras, "Robust Recovery of the Epipolar Geometry for an Uncalibrated Stereo Rig", *European Conference on computer Vision*, Vol.1, pp. 567-576, May 1994
11. Bill Triggs, "Factorization methods for projective structure and motion", *Proc. of Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Soc*. Press, pp. 845-851, 1996
12. Kiriakos N. Kutulakos, and James R. Vallino, "Calibration-Free Augmented Reality" *IEEE Transaction of Visualization and Computer Graphics*, vol.4, no.1, 1998
13. A.Chiuso, P.Favaro, H.Jin, and S.Soatto, "Structure from motion causally integrated over time pattern Analysis and Machine Intelligence", *IEEE Transactions on*, Vol. 24 Issue:4, pp. 523 -535, April 2002

3-D Pose Tracking of the Car Occupant

Sang-Jun Kim[1,2], Yong-Guk Kim[3], Jeong-Eom Lee[1],
Min-Soo Jang[1], Seok-Joo Lee[4], and Gwi-Tae Park[1]

[1] Dept. of Electrical Engineering, Korea University, Seoul, Korea
gtpark@korea.ac.kr
[2] Interdisciplinary Programs of Mechatronics, Korea University, Seoul, Korea
[3] School of Computer Engineering, Sejong University, Seoul, Korea
ykim@sejong.ac.kr
[4] Hyundai Autonet Co. Ltd., Korea

Abstract. Although airbags in the car play an important role for the safety of occupants, in fact, many peoples have been injured or killed by the deployment of airbags themselves. Such conventional airbags are deployed by the shock sensors. As an alternative approach, a vision-based smart airbag system could be promising. This paper describes a new method by which 3-D pose of the car occupant can be recognized. We combine 2-D head tracking information with a disparity map of the occupant for 3-D pose tracking. Result shows that the system can locate the head position around the passenger's seat with a real-time basis.

1 Introduction

Airbags in the car play an important role for preventing life-threatening head injuries by avoiding direct impact to the dashboard during the accident. Although airbags have saved many lives, however, many occupants, in fact, have been killed by the deployment of the airbags themselves, according to the National Highway Traffic Safety Administration (NHTSA) report [1]. To reduce the risk caused by airbag, it is important to design a smart airbag that could control intensity of its deployment. In such case, it is necessary to recognize occupant's pose before triggering the airbag. Normally, the position of the occupant's head provides important information for recognizing the occupant's pose. Recently, several researchers have studied how to detect the head of the occupant [2, 3, 4, 5].

Movements of the car occupant can be occurred by moving his body or by car movement. Often, such movements can be measured by comparing two sequential images, captured by a camera installed within the car. However, in the present study, we would like to track the head of the occupant using the region tracking method. As the position of the tracked head provides only 2-D information, we combine the head tracking with 3-D disparity information extracted from the stereo images to acquire 3-D pose of the occupant.

The present paper consists of several sections as follow. The image tracking method is discussed in section 2. Pose recognition system is described in section 3. Result of experiments is reported in section 4. Finally, our result is summarized, and the performance of the whole system is discussed in section 5.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 219–224, 2005.

2 Head Image Tracking

Researchers have been working on tracking of the object for several decades. One of the popular methods was proposed by Lucas-Kanade [6]. And yet it is rather slow in applying that method to the real application such as the head tracking problem. Recently, reformulation of the original version, known as the *inverse compositional* image alignment method, can be fast and reliable [7]. Basically, the goal of Lucas-Kanade algorithm is to minimize the sum of squared error between two images, namely, the template T and the input image I :

$$\sum_{\mathbf{x}}[I(\mathbf{W}(\mathbf{x};\mathbf{p}))-T(\mathbf{x})]^2 \tag{1}$$

where $\mathbf{x}=(x,y)^T$ containing the pixel coordinates; $\mathbf{W}(\mathbf{x};\mathbf{p})$ is the parameterized set of allowed warps; $\mathbf{p}$ is a vector of parameters. Since the tracking object moves in 3-D, we should consider the set of affine warps. So, the warps $\mathbf{W}(\mathbf{x};\mathbf{p})$ can be expressed as follow:

$$\mathbf{W}(\mathbf{x};\mathbf{p})=\begin{pmatrix}1+p_1 & p_3\\ p_2 & 1+p_4\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix}+\begin{pmatrix}p_5\\ p_6\end{pmatrix} \tag{2}$$

where $\mathbf{p}$ consist of 6 parameters, so $\mathbf{p}=(p_1,p_2,\ldots,p_6)^T$.

Warping I back to computing $I(\mathbf{W}(\mathbf{x};\mathbf{p}))$ requires interpolation of the image I at the sub-pixel locations $\mathbf{W}(\mathbf{x};\mathbf{p})$. To optimize the parameter $\mathbf{p}$, Lucas-Kanade algorithm assumes that the current estimated $\mathbf{p}$ is known, and then it intends to solve the equation (1) by incrementing parameters $\Delta\mathbf{p}$ iteratively as shown by the following expression:

$$\sum_{\mathbf{x}}[I(\mathbf{W}(\mathbf{x};\mathbf{p}+\Delta\mathbf{p}))-T(\mathbf{x})]^2 \tag{3}$$

by updating the parameter.

$$\mathbf{p}\leftarrow\mathbf{p}+\Delta\mathbf{p} \tag{4}$$

These steps are iterated until the parameter $\mathbf{p}$ converges. This updating rule is referred as the additive approach. The main idea of the 'inverse compositional algorithm' is to reverse the order between the template and the input image as shown in the following expression:

$$\sum_{\mathbf{x}}[T(\mathbf{W}(\mathbf{x};\Delta\mathbf{p}))-I(\mathbf{W}(\mathbf{x};\mathbf{p}))]^2 \tag{5}$$

with respect to $\Delta\mathbf{p}$. And it updates the estimate of the warp as follow:

$$\mathbf{W}(\mathbf{x};\mathbf{p})\leftarrow\mathbf{W}(\mathbf{W}(\mathbf{x},\mathbf{p})^{-1};\mathbf{p}) \tag{6}$$

As equation (5) is non-linear, its first-order Taylor expansion is given by:

$$\sum_{\mathbf{x}}\left[T(\mathbf{W}(\mathbf{x};\mathbf{0}))+\nabla T\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\Delta\mathbf{p}-I(\mathbf{W}(\mathbf{x};\mathbf{p}))\right]^2 \tag{7}$$

Assuming again that $\mathbf{W}(\mathbf{x};\mathbf{0})$ is the identity warp, the solution to this least-squares problem is given as follow.

$$\Delta \mathbf{p} = H^{-1}\sum_{\mathbf{x}}\left[\nabla T\frac{\partial \mathbf{W}}{\partial \mathbf{p}}\right]^T\left[I(\mathbf{W}(\mathbf{x};\mathbf{p})-T(\mathbf{x})\right] \tag{8}$$

Where H is Hessian matrix and I is replaced by T :

$$H = \sum_{\mathbf{x}}\left[\nabla T\frac{\partial \mathbf{W}}{\partial \mathbf{p}}\right]^T\left[\nabla T\frac{\partial \mathbf{W}}{\partial \mathbf{p}}\right] \tag{9}$$

In the inverse compositional algorithm, since we can calculate Hessian matrix and other parameters before tracking, it allows the fast iteration and leads to the real-time basis tracking of the object. Fig. 1 illustrates the sequence of the head tracking method.

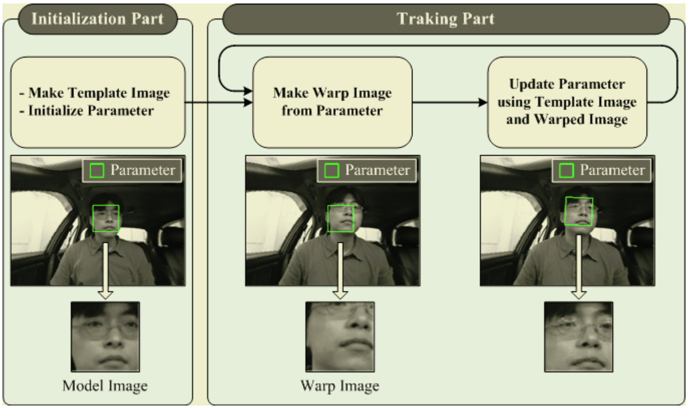


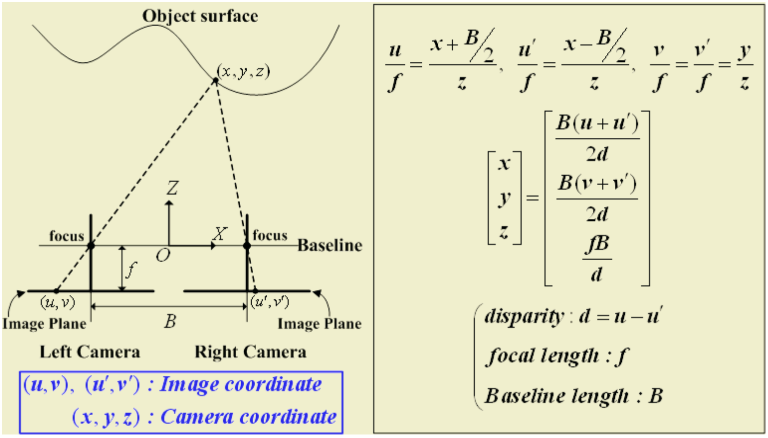
Fig. 1. The sequence of the head tracking method

In the initialization step, a template image is cropped from an input image, and then warping parameter is initialized and Hessian matrix is calculated. In the tracking step, a warped image is cropped by the warping parameter, and then the parameter is updated by inverse compositional algorithm using the template and warped images.

3 Stereo Vision for 3-D Pose Information of the Car Occupant

Although the head tracking method give us useful information on the present head position of the occupant, it does not provide 3-D pose information of him since the tracking is carried out only on 2-D image. In the present study, we aim to combine the head tracking with stereo information to get the pose information of the occupant, sitting in the car seat. The stereo cameras installed in the car provide 3-D information of the occupant as illustrated in Fig. 2.

An object seen from a pair of stereo cameras leads to a visual difference, which depends on the distance B between two cameras as well as 3-D shape of the object as depicted in Fig. 2(a). Fig. 2(b) shows that such image difference between two stereo images, called the disparity map, can be calculated using a stereo algorithm [8]. Since our goal is to compute the occupant's 3-D pose, we transform the position of the head on the image coordinates acquired by head tracking into a 3-D coordinate by combining the disparity map. Indeed, such 3-D coordinates can be presented on a top-view graph as illustrated in Fig. 3, in which each segmented-area designates the location the head belongs to at a certain moment. In a real-time demo, the head indicated as a circle drifts around different areas as the occupant move his head and body around the seat.



(a)

$$\frac{u}{f}=\frac{x+B/2}{z},\ \frac{u'}{f}=\frac{x-B/2}{z},\ \frac{v}{f}=\frac{v'}{f}=\frac{y}{z}$$

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}=\begin{bmatrix} \frac{B(u+u')}{2d} \\ \frac{B(v+v')}{2d} \\ \frac{fB}{d} \end{bmatrix}$$

$$\begin{cases} disparity : d = u - u' \\ focal\ length : f \\ Baseline\ length : B \end{cases}$$

(b)

Fig. 2. (a) Schematic diagram of the stereo vision and (b) 3-D position of the object surface (b)

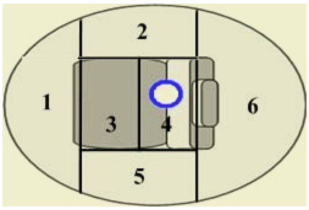


Fig. 3. Six areas where the occupant's head can be located

4 Results

4.1 Database and Behavior Patterns

A set of video database is recorded within an experimental car with a 15 fps (frames per second) speed for three subjects as shown in

Table. 1. The database consists of two kinds of behavior patterns: one is 'simple behavior pattern' where the occupant only moves his head and shoulder as shown in Fig. 4(a); the other is 'complex behavior pattern' where the subject allows to move his arm as well, and so his arm partly occludes his head as shown in Fig. 4(b). An extra video database is also recorded for two subjects within an infrared illumination in the dark situation, since we will add it for the night drive test.

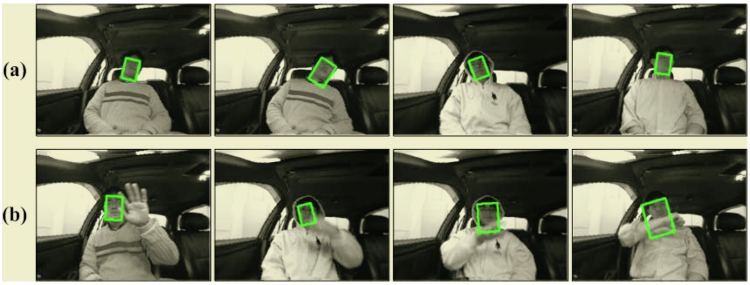


Fig. 4. Head tracking for two behavior patterns (a) The simple pattern, (b) the complex pattern

Table 1. The video database (unit: frame)

	Behavior Pattern		Remark
	Simple	Complex	
Occupant Ⅰ	37	40	Infrared: 47
Occupant Ⅱ	51	48	Infrared: 35
Occupant Ⅲ	40	38	•
Total	128	126	Infrared: 82

4.2 Experiments and Result

The experiments were carried out with an off-line basis for three different video databases. We only counted it as a correct image frame when the system is able to assign the occupant's head position to the ground-truth among those 6 areas (see Fig. 3). Since the tolerance of the error was 7.5 cm, any image frames having an error larger than this tolerance are counted as the incorrect frames as shown in Fig. 5, where the error is varied as the frame goes by time. Table. 2 summarizes the result of the analysis. It shows that the success rate for the simple patterns was 90.0%, and that for both patterns was 83.3%.

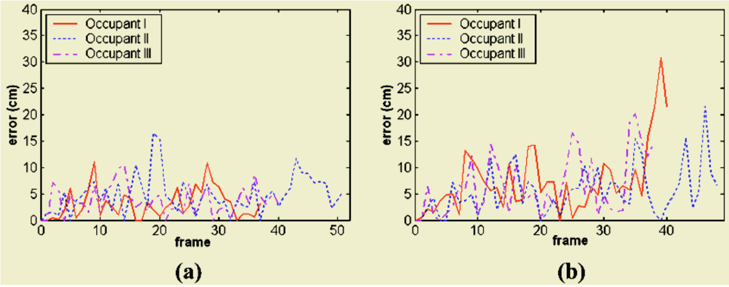


Fig. 5. Graphs for the trial errors; (a) simple pattern images and (b)complex pattern images.

Table 2. Experimental Result (unit: frame)

pattern	video frames	correct frames	incorrect frames
simple	128	117	11
complex	126	91	35
infrared	82	72	10
total	336	280	56

5 Conclusions and Future Work

We propose a new method, which locates and tracks the occupant's head for a smart airbag triggering system. The main idea of this method is to combine the head tracking with stereo information to acquire 3-D pose of the occupant in real-time basis. For the head tracking, we have adopted a recent Lucas-Kanade tracker, called the inverse compositional algorithm. The disparity map for the occupant is calculated from the stereo image captured using a stereo camera. We have tested the system for diverse behavior patterns of the occupant including a night drive case. Result suggests that the new method tracks the occupant reasonably well. We plan to combine the present tracker with a head detector which is based on motion information of the occupant for the reliable tracking to supplement the missing frames with the present method.

Acknowledgments

This work was supported by Hyundai Autonet Co. and the 21st century Frontier Program.

References

1. National Highway Traffic Safety Administration. Air Bag Fatal and Serious Injury Summary Reports. http://www-nrd.nhtsa.dot.gov/departments/nrd-30/ncsa/TextVer/SCI.html
2. Y. Owechkp, N. Srinivasa, S. Medasani, and R. Boscolo, "Vision-Based Fusion System for Smart Airbag Applicaions", IEEE, Intelligent Vehicle Symposium, vol. 1, pp. 245-250, 2002
3. R. Reyna, A. Giralt, and D. Esteve, "Head Detection Inside Vehicles with a Modified SVM for Safer Airbags", IEEE Intelligent Transportation Systems Conference, pp. 268-272, 2001
4. B. Alefs, M. Clabian, H. Bischof, W. Kropatsh, and F. Khairallah, "Robust Occupancy Detection from Stereo Images", IEEE Intelligent Transportation Systems Conference, 2004
5. K. Huang and M. Trivedi, "Robust Real-Time Detection, Tracking, and Pose Estimation of Faces in Video Streams" 17th International Conference on Pattern Recognition, vol. 3, pp. 965-968, 2004
6. B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674-679, 1981.
7. S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework", *International Journal of Computer Vision*, 56(3):221-225, Feb. 2004. Previously appeared as CMU Robotics Institute Technical Report CMU-RI-TR-02-16.
8. C. Sun, "A Fast Stereo Matching Method", Digital Image Computing: Techniques and Applications, pp.95-100, Massey University, Auckland, New Zealand, December 10-12, 1997.
9. R. Hartley and A. Zisserman, "Multiple View Geometry in computer vision", Cambridge University Press, 2000.
10. G. Baxes, "Digital Image Processing: principles and application*s*", John Wiley & Sons, New York, USA, 1994.

Face Recognition by Multiple Classifiers, a Divide-and-Conquer Approach

Reza Ebrahimpour[1,2], Saeed Reza Ehteram[3], and Ehsanollah Kabir[4]

[1] School of Cognitive Sciences, Institute for Studies on Theoretical Physics and Mathematics,
Niavaran, Tehran, Iran
[2] Department of Electrical Engineering,
Shahid Rajaee Teachers Training University, Tehran, Iran
ebrahimpour@ipm.ir
[3] Department of Electrical Engineering, Pars Electric High Education Center,
Jamee University of Applied, Science and Technology, Farahzad, Tehran,
P.O.Box 16356-83863, Iran
sa_ehteram@yahoo.com
[4] Department of Electrical Engineering, Tarbiat Modarres University,
Tehran, P.O. Box 14115-143, Iran
kabir@modares.ac.ir

Abstract. In this paper, an approach that uses a combination of neural network classifiers (CNNC) is applied to human face recognition. We present a divide-and-conquer approach for system composed of several separate networks. Decomposing the complex problem into sub-problems for solving them by a binary base classifier is presented. Each of that learns to recognize a subject of the complete set of training database. Combining the results of sub-problems with max rule accomplished to achieve better performance. The recognition rate of 100% for ORL and Yale database was obtained using the mentioned devised algorithm.

Keywords: Face Recognition, Neural Networks, Classifier Combination, Multilayer perceptron, Decomposition

1 Introduction

In the last years, Face Recognition has become one of the most challenging tasks in the field of pattern recognition. The recognition of faces is very important for many applications: video-surveillance, retrieval of an identity from a database for criminal investigations and forensic applications [1].

The face is considered a good biometric for many reasons: the acquisition process is non-intrusive. The acquisition process of a face from a scene is simpler and cheaper than the acquisition of other biometrics as the iris and the fingerprint. On the other hand, many problems arise, because of the variability of many parameters: face expression, pose, scale, lighting, and other environmental parameters. For this reason, we can subdivide the applications, which involve face recognition in two categories: applications in a controlled environment and applications in an uncontrolled environment. The first kind of applications refers to the problem of "identity authentication": A subject submits to the system its face (frontal and/or profile view) and he declares

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 225–232, 2005.

his identity. The aim of the system is to verify the matching between the claimed identity and the given biometric. This kind of application is typical for Internet transactions, driver’s licenses, and access to limited areas. The second kind of applications refers to the problem of “recognition of an identity in a scene”, and it is typical for video-surveillance applications. A system that automatically recognizes a face in a scene first detects it and normalizes it with respect to the pose, lighting and scale. Then, the system tries to associate the face to one or more faces stored in its database, and gives the set of faces that are considered as “nearest” to the detected face. This problem is much more complex than the “verification” problem, and it requires more computational resources and very robust algorithms for detection, normalization and recognition. Usually, each of these problems is so complex that it must be studied separately.

Many real-world problems are too large and too complex for a single monolithic system to solve alone. There are many examples from both natural and artificial systems, which show that a composite system consisting of several subsystems can reduce the total complexity of the system while solving a difficult problem satisfactorily. The success of neural network ensembles in improving classifier's generalization is a typical example [2].

Given the advantages of multiple classifiers and the complexity of the problems that are beginning to be investigated, it is clear that multiple classifiers processing is and will be an important and pervasive problem-solving technique. Multiple classifiers design in engineering, however, has relied on human expertise (often a committee) to manually divide a system into specialized parts, often in an ad hoc manner. While manual design may be appropriate when there are experienced human experts with sufficient prior knowledge of the problem to be solved, it is certainly not the case for those real-world problems about which we do not have much prior knowledge. Tedious trial-and-error processes are often involved in designing multiple classifiers in practice.

The idea of combining several techniques has received considerable attention and it has been used in different applications. Kittler et al [3] present an elementary theoretical framework, which can be used to generate some simple combination rules when adopting different assumptions and different styles of expression, and a study on different strategies for combining the classifiers, was presented by Kuncheva [4]. Nevertheless, research is still needed to solve the problem of how to create an appropriate criterion for combining the classifiers.

Recently, a pattern recognition method based on combination of neural networks has been proposed. In literature, Caleanu [5], the Neural Networks Committee Machines is employed to implement facial recognition. The original images are divided into several blocks, each of which is learned by a neural network module. Jing et al. [6] an approach that uses a combination of linear classifiers is applied to face recognition. A genetic algorithm is used to generate the rational weights for the classifiers.

In this article, a multi class problem decomposed to several binary problems and then results of binary classifiers composed with max rule. The rest of this paper is organized as follows: Section 2 describes multiple classification in detail and gives motivations and ideas behind various design choices; Section 3 presents experimental results on multiple classification and some discussions; and finally Section 4 concludes with a summary of the paper and a few remarks.

2 Multiple Classifications

If back propagation is used to train a single, multilayer network to perform different sub-tasks on different occasions, there will generally strong interface effects that lead to slow learning and poor generalization. In combining, the main requirement is to have some individual classifier that not only has an acceptable performance but also exhibits independence in decision making. Generally speaking, the parts of error that are more likely removed are not common to all components. This is why correlation reduction between classifiers is important, as explained in continuing.

Since classifiers are made through a training procedure, to have classifiers, which generalize diversely, they should be trained differently. The training procedure can be affected by input representation of patterns, training samples, learning procedure, and supervision strategy, on which correlation reduction techniques will be based.

It is assumed that a simple MLP neural network composed of a single hidden layer and the output layer is capable of solving difficult and complex problems like the non-linear Systems problem. Non-linear hyperbolic tangent activation functions may be used for the hidden units while linear activation functions may be used in the output units. It must be noted that feed-forward MLP is the most extended Artificial Neural Networks architecture in engineering and pattern recognition Systems applications.

In classification a given pattern x should be assigned to one of several possible classes. Based on this fact, in supervised learning, pairs of input desired output are used. So changing desired target could change the environment. If we redefine the classes for example by giving the same label to a group of classes, the classifiers that are made by different defining of classes, can generalize diversely. Methods like ECOC [7], binary clustering [8], and pair wise coupling [9] used in method. Binary classification interests us because some accurate and efficient algorithms for solving 2-class classification problems don't naturally scale up to multi-class. Furthermore some useful algorithms that make the classifiers more capable are designed for binary classification. Although some attempts were made to scale them up their performance on binary cases was better. Binary classifiers have some advantage in software and hardware implementation because of their similarity with binary logics.

According to above descriptions binary classification in CNNC is used as an expert in composite system. In training phase of CNNC, each expert is informed to learn a singular class. Therefore in construction of classifiers combination used in composite system, classifiers are maturated uncorrelated.

Therefore, Let $D = \{D_1,...,D_C\}$ be the classifier ensemble and $\Omega = \{w_1,...,w_C\}$ be the set of class labels. The individual output is estimate of the class score, that is, the output $d_i(x)$ of classifier D_i in support of the hypothesis that x comes from class w_i and no from other classes, is an estimate of $p(w_i|x), i = 1,...,C$. Therefore after training each classifier independently, the unknown input data presented for each classifier. The result of classifiers as a vector with C rows, as an input for combiner, presented. Finally, the output of combiner determines the class label of the input data. In this method, max rule as a combiner strategy is used.

3 Experimental Studies

We have applied CNNC to two benchmark problems, including the ORL face database and Yale face database. The ORL face database contains a set of faces taken between April 1992 and April 1994 at the Olivetti Research Laboratory in Cambridge, UK1. There are 10 different images of 40 distinct subjects. For some of the subjects, the images were taken at different times. There are variations in facial expressions (open/closed eyes, smiling/non-smiling), and facial details (glasses/no glasses). All the images were taken against a dark homogeneous background with the subjects in an up-right, frontal position, with tolerance for some tilting and rotation of up to about 20 degrees. There is some variation in scale of up to about 10%. and a larger set of images for one subject is shown in Fig.1. The images are grayscale with a resolution of 92×112 pixels as frontal views.

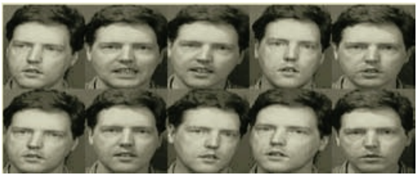


Fig. 1. The set of 10 images for one subject from ORL face database

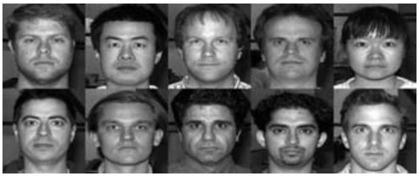


Fig. 2. Samples of face images in the yale face database

The Yale face database contains 165 face images of 15 individuals. There are 11 images per subject, one for each facial expression or configuration: center-light, glasses/no glasses, happy, normal, left-light, right-light, sad, sleepy, surprised and wink. Samples of Yale face database are shown in Fig.2

In this paper, we have used PCA [1] as a feature domain that uses global data to create the feature vector elements. However, to keep only the important data about the face images, and to eliminate the irrelevant data the feature extraction is done in two steps. In the first step, by using the shape information based on [10], we have created a pure-face to enclose only the important information needed for the recognition algorithm, while in the second step, the feature vector has been obtained through calculation of the PCA of the derived pure-face.

3.1 Pure-Face Formation

The pure-face encloses all the pertinent information around the face in an ellipse while the pixel value outside the ellipse is set to zero. A technique is presented in Refs. [10], which finds the best-fit ellipse to enclose the facial region of the human face in a frontal view of the facial image. Unfortunately through the creation of the pure-face with the best-fit ellipse many unwanted regions of the face image may still appear in this pure-face. These include hair portion, neck and part of the background as an example. Instead of using the best-fit ellipse for creating a pure-face we have defined another ellipse. The new ellipse has the same orientation and center as the best-fit ellipse but the lengths of its major and minor axes are calculated from the length of the major and minor axes of the best-fit ellipse.

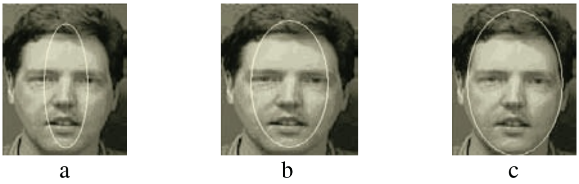


Fig. 3. Pure-face formation based on different values of ellipse parameters, part c is better than a, b for face processing

3.2 Principle Component Analysis (PCA)

PCA is a well-known statistical technique for feature extraction. Each $M \times N$ image in the training set was row concatenated to form $MN \times 1$ vector x_k. Given a set of training images $\{x_k\}_{k=0,1,\ldots,N_T}$ the mean vector of the training set was obtained as [1].

$$\bar{x} = \frac{1}{N_T}\sum_{k=1}^{N_T} x_k \tag{1}$$

A $N_T \times MN$ training set matrix $X = [x_k - \bar{x}]$ can now be built. The basis vectors are obtained by solving the eigenvalue problem:

$$\lambda = V^T \textstyle\sum_X V \tag{2}$$

Where $\sum_X = XX^T$ is the covariance matrix, V is the eigenvector matrix of $\sum_x$ and λ is the corresponding diagonal matrix of eigenvalues. As the PCA has the property of packing the greatest energy into the least number of principal components, eigenvectors corresponding to the m largest eigenvalues in the PCA are selected to form a lower-dimensional subspace. It is proven that the residual reconstruction error generated by discarding the $N_T - m$ components is low even for small m.

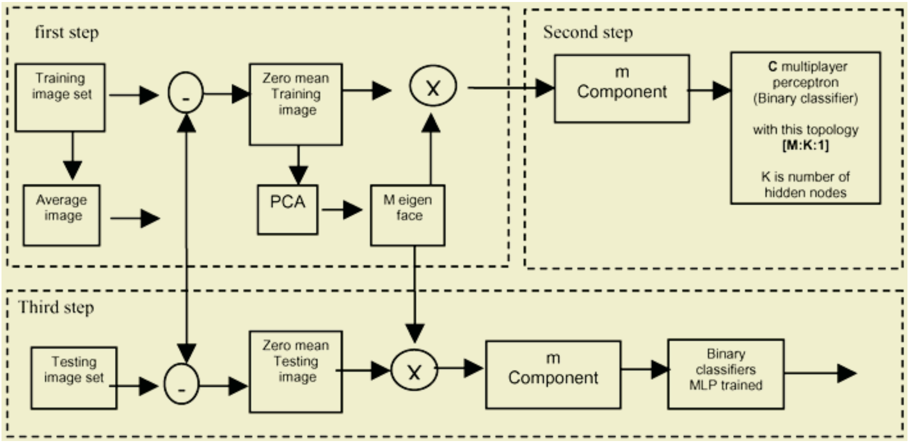
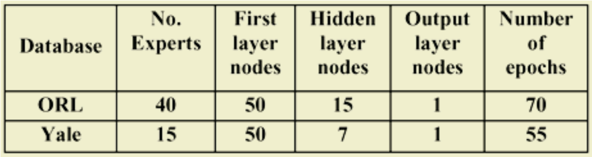


Fig. 4. Blok diagram of CNNC system

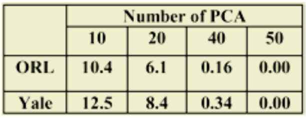
Fig.4 illustrates the experiments carried out in this work. Each experiment consists of three steps: generation of the feature vector, training the classifiers and testing the classifiers. In the first step, The training and testing set is selected, by randomly choosing five images for each subject from the ORL database and six images from the Yale database. Therefore, in the ORL database a total of 200 images are used as the training set and another 200 are used as the testing set while in the Yale database a total of 90 images are used for training and the rest are used for testing. Then PCAs are generated inside the pure-faces. In the second step, the classifier is designed and trained. Finally in the Third step, the performance of the classification is evaluated. This procedure has been repeated for each learning algorithm by randomly choosing different training and testing sets. According to ultimate accomplish each expert is trained by properties that listed in Table1.

Table 1.



Database	No. Experts	First layer nodes	Hidden layer nodes	Output layer nodes	Number of epochs
ORL	40	50	15	1	70
Yale	15	50	7	1	55

Table 2.



	Number of PCA			
	10	20	40	50
ORL	10.4	6.1	0.16	0.00
Yale	12.5	8.4	0.34	0.00

3.3 Experimental Results

Table 2 show the error rate computed with 5 times repeat for the CNNC as a function of the number of the PCA for the ORL and Yale databases.

Comparisons between Other Work Direct comparison and other decomposition approaches to designing ensembles are very difficult due to the lack of such results. Instead, the best and latest results available in the literature were used in the comparison. Table 3 show the outcome of this comparative study, where SINN denotes the shape information with the neural network that was reported in Ref. [11]. CNN is the convolution neural network method in Ref. [12]; NFL is the nearest feature line method in Ref. [13], and FT denotes the fractal transformation technique in Ref. [14], FHLA stands for fuzzy hybrid learning algorithm in Ref.[8], and Ensemble Neural Networks with Co-Evolutionary Algorithm in Ref. [15]. In this table, the CNNC yielded an error rate of 0.00%, which is the best obtained in our experiment for the ORL database.

Table 3. Error rate for different human face recognition system on the ORL database

Algo-rithm	CNN	NFL	FT	SINN	FHLA	CELS	CNNC
Error Rate(%)	3.83	3.125	1.75	1.323	0.45	0.39	0.00

Each of which classes in this paper are constructed by 50 input nods (constant value). Hidden nods and number of epochs would be change referring to hardness of each class ,as is shown below out put layer consist of just one node because we have 40 classes .these two tables below are shown this fact that classes which are belong to human faces with sunglasses on one of their images or smiling energetically or have critical differences between their images are belonged to hard classes(differences between 8 image using for training phase). And these classes need more epochs and more hidden layers as is ordered below:

Table 4.

Classes	Hidden layer	Epochs	Output nods	Ultimate percent For ORL
1 up to 16	8	10	1	100%
23 up to 27	12	10	1	
17 up to 22	15	10	1	
31 up to 37	7	10	1	
28 up to 31	21	10	1	
38 up to 40	11	10	1	

Table 5.

Classes	Hidden layer	Epochs	0Output nods	Ultimate percent For ORL
1 up to 16	10	7	1	100%
23 up to 27	10	15	1	
17 up to 22	10	23	1	
31 up to 37	10	6	1	
28 up to 31	10	42	1	
38 up to 40	10	12	1	

4 Conclusions

CNNC provides a simple way of designing Neural Networks ensembles, where each Neural Networks is an individual or a representative from each species in the population. The binary classification and max rule combination was adopted to encourage the formation of species in the population. The proposed method offers faster training, less bulk of training and computing and less epochs used to train networks in comparing with other methods. Comparison with some of the existing traditional technique in the literatures on the same databases indicates the usefulness of the proposed technique. The recognition rate of 100% for ORL and Yale database was obtained using the mentioned devised algorithm.

References

1. Turk, M., Pentland, A.: Eigenfaces for Recognition, Journal of Cognitive Neuroscience, (3)(1991) 71-86.
2. Ghaderi, R.: Arranging simple Neural Networks to solve Complex classification problems, Ph.D. Thesis, Center for Vision, speech and signal processing, university of Surry, (2000).
3. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifers, IEEE Trans. Pattern Anal. Mach. Intel. 20 (3) (1998) 226–239.
4. Kuncheva, L.I.: A theoretical study on six classi!er fusion strategies, IEEE Trans. Pattern Anal. Mach. Intel. 24 (2)(2002) 281–286.
5. Caleanu, C. D.: facial recognition using committee of neural networks, 5th seminar on Neural Network Applications in Electrical Engineering, NEUREL- (2000).

6. Jing, X., Zhang, D.: Face recognition based on linear classifiers combination, Neurocomputing, Vol. 50, (2003) pp. 485-488.
7. Dietterich, T.G., Bakiri, G.: Solvig multi-class learning problems via error-correcting output codes, Journal of Artificial Intelligence Research, Vol. 2, (1995) pp. 263-286.
8. Wilson, C.L., Grother, P.J., Barnes, C.S.: Binary decision clustering for neural network based optical character recognition, pattern recognition, vol. 29, No. 3, (1996) pp. 425-437.
9. Hastie, T., Tibshirani, R.: Classification by pairwise coupling, technical report 94305, development of statistics, Stanford University, (1996).
10. Haddadnia, J.,Faez, K.,Ahmadi, M.: A fuzzy hybrid learning algorithm for radial basis function neural network with application in human face recognition, Pattern Recognition 36(5): (2003)1187-1202.
11. Haddadnia, J., Faez, K.: Human face recognition with moments invariant, Proceeding of The IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing, Maryland, USA, June 3–6, (2001).
12. Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D.: Face recognition: a convolutional neural networks approach, IEEE Trans. Neural Networks (Special Issue on Neural Networks and Pattern Recognition) 8 (1) (1997) 98–113.
13. Li, S.Z., Lu, J.: Face recognition using the nearest feature line method, IEEE Trans. Neural Networks 10 (1999) 439–443.
14. Tan, T., Yan, H.: Face recognition by fractal transformations. IEEE Int. Conf. Acoustics, Speech Signal Process. 6,(1999) pp. 3537–3540.
15. Mazloom, M., Ebrahimpour, R., Lucas, C.: Face Recognition: An Ensemble Neural Networks With CO-Evolutionary Algorithm Approach", In the 2nd IEEE Conference on Advancing Technology in the GCC: Challenges, and Solutions, November 23- 5, (2004).

Shape Comparison of the Hippocampus Using a Multiresolution Representation and ICP Normalization

Jeong-Sik Kim[1], Yong-Guk Kim[1], Soo-Mi Choi[1], and Myoung-Hee Kim[2]

[1] School of Computer Engineering, Sejong University, Seoul, Korea
smchoi@sejong.ac.kr
[2] Dept. of Computer Science and Engineering, Ewha Womans University, Seoul, Korea
mhkim@ewha.ac.kr

Abstract. It is known that deformation of the hippocampus shape is involved with several neurological diseases. In this paper, we propose a hybrid shape representation scheme, which consists of multiresolution skeletons, voxels and meshes for the shape analysis of the hippocampus. Initially, a hippocampal surface model is reconstructed from MRI and then it is placed into a canonical coordinate system, where the position, orientation and scaling are normalized. From the voxel representation of the hippocampus, multiresolution skeletons are extracted and Iterative Closest Point normalization is carried out. Then the shape similarity of two hippocampal models is computed with a hierarchical fashion. In addition, we have implemented a neural network based classifier to discriminate whether a hippocampal model is normal or not. Results indicate that the proposed hybrid representation and the skeleton-based normalization using ICP are very effective in 3D shape analysis of the hippocampus.

1 Introduction

Analyzing the hippocampal structure in the brain is an important issue in medical area. In particular, it is known that an abnormal shape of the hippocampus is involved with neurological diseases such as epilepsy, schizophrenia, and Alzheimer's diseases [1]. Many researchers have been tried to investigate the shape change pattern of hippocampus related to these diseases. Therefore, in order to estimate the shape difference reliably, it is essential to select an efficient shape representation scheme and to apply the robust shape normalization approach to this representation. The scale and orientation of the hippocampus segmented from the MRI vary depending on patient's age and sex, etc. So, we need to normalize the pose of the hippocampus model to obtain an accurate similarity measurement. A typical tool for accomplishing the pose normalization has been the Principal Component Analysis (PCA) method [2-4]. However, it often does not guarantee a reliable result for some models, having extensive local shape deformation.

A new method adopting the multiresolution representation and the Iterative Closest Points (ICP) normalization process allows us to analyze 3D shape by changing the level-of-detail, and moreover it is possible to increases the speed of shape comparison without degrading accuracy. In this paper, in order to provide a reliable shape comparison, we adopt ICP normalization rather than the conventional PCA method and implement a neural network classifier for discriminating between normal controls and epileptic patients.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 233–239, 2005.

The rest of this paper is organized as follows. Section 2 describes the multiresolution shape representation and Section 3 describes the ICP-based normalization using skeletons. Section 4 explains the global and local shape analysis and the implementation of a classifier. Experimental results and discussion are given in Section 5 and some conclusions and future works are given in Section 6.

2 Octree-Based Multiresolution Shape Representation

In this section, we describe how to represent the shape of hippocampal structure using a hierarchical LOD approach. Initially, we segment the hippocampal structure from the MRI of the brain, and multi-level surface meshes are generated using the Marching Cube algorithm [5]. Then, we convert the polygonal surface to an intermediate binary voxel representation using a depth-buffer voxelization, which makes it easier to extract a skeleton as well as to relate to the original medical images. The extracted skeleton is used for sampling the meshes, computing a similarity measure between the shapes, and placing the shape into a canonical coordinate system. As we separate and store these representations using the Octree structure, it is possible to reduce the computation time for the similarity estimation, and to capture the local shape difference with a hierarchical fashion. The Octree is a data structure to represent objects in 3D space, automatically grouping them hierarchically and avoiding the representation of empty portion of the space. Three different types of shape information (i.e. meshes, voxels, and skeletons) are integrated into an Octree data structure. Fig. 1 illustrates how to integrate these data, where the shape information in the different nodes is labeled with different colors.

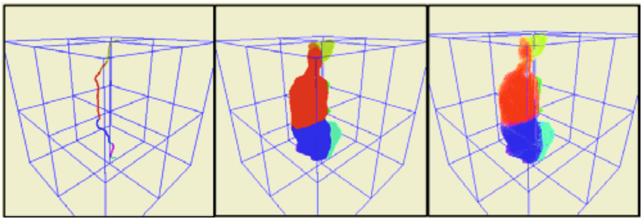


Fig. 1. Octree based hybrid representation of the hippocampus shape: (a) skeletal model; (b) mesh model; (c) voxel model

3 ICP-Based Normalization for Shape Comparison

Besl *et al.* [6] introduce the ICP algorithm in registering 3D primitives. It is an optimization approach by minimizing the least-square errors during the registration stage with an iterative way. We adopt the ICP algorithm proposed by Zhang *et al.* [7], in which they apply a free-form curve representation to the registration process. Our method consists of three steps. First, we extract 3D skeletons from two shapes and then compute the optimal transformation matrices to minimize the least-squares between two shape models. Secondly, these matrices are applied to the initial model, and finally the transformation is ended when the errors are converged to user-defined threshold. At this time, we define the first skeleton points set as a “Model” and the second set as a “Shape” The “Model” is a skeleton that does not change. And the “Shape” is transformed. Fig. 2 shows the algorithm for normalizing two models.

```
1. Input the fixed points set $F=\{f_i\}$, the moving points set $M=\{m_i\}$ $(i=1,...,N_k)$.
2. Initialize: (1) $D_{\max}$ = larger $D$, (2) $M_0=M, H=I$ (Identity transformation), $k$=0.
3. Iterate until it converges.
    1) Find the closest points: $Z_k= C(M_k, F)$
    2) Compute the pseudo point matching
    3) Update the matching
    4) Compute the difference: $(H_k,d_k) = E(M_0,Z_k)$
    5) Apply the motion to all points: $Y_{k+1}=H_k(M_0)$
    6) Terminate when $d_k-d_{k+1}$ < threshold
```

Fig. 2. The ICP algorithm for normalizing using two skeletons

Given two skeletons, $C_i(i=1,\dots,m)$, and $C'_k(k=1,\dots,n)$ extracted from two models, in order to normalize these objects, we adopt an object function to find the optimal translation t and rotation R defining the transformation corresponding the “Shape” skeleton to the “Model” skeleton. Eq. (1) defines the object function. $d(x,C)$ is the distance from a point x to a skeleton C. If $x_{i,j}$ has the closest point for the skeleton C', the value of $p_{i,j}$ is 1. If not, that value is 0. ($q_{k,l}$ analogously). The final result of the optimal normalization is decided when function F is minimized. The function F converges into the minimum value when the result of Eq. (2) is the maximum value. Eq. (3) chooses the closest point from a pair of the 3D skeleton sets.

$$F(R,t)=\sum_{i=1}^{m}\sum_{j=1}^{N_i} p_{i,j}\,d^2(Rx_{i,j}+tC'_k)+\sum_{k=1}^{n}\sum_{l=1}^{N_k} q_{k,l}\,d^2(R^T x'_{k,l}-R^T t, C_i) \tag{1}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{N_i} p_{i,j}+\sum_{k=1}^{n}\sum_{l=1}^{N_k} q_{k,l} \tag{2}$$

$$d(x,C'_k)=\min_{l\in\{1,\dots,N_k\}} d(x,x'_{k,l}) \tag{3}$$

The pseudo point-matching is a method for removing wrong pairs of closest points as we compare the statistical values to the distances of the all pairs of closest points. The first constraint $D_{\max}$ is the maximum acceptance value. And the constraint angle θ ensures the robustness of an orientation. For the “Model” and “Shape” models, the distance measured from these two models is $\{x_i\}$ and $\{y_i\}$, respectively, and $\{d_i\}$ is the deviation from the mean value. Then we compute pseudo point matching pairs by changing $D_{\max}$ according to the intermediate result of the registration flexibly. Eq. (4) defines a mean value μ and a deviation value σ. In Eq. (4), when the registration result is bad, n is decreased as one by one. Finally, when n is 1, $D_{\max}$ has the best registration status.

$$\mu=\frac{1}{N}\sum_{i=1}^{N} d_i,\ \sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i-\mu)^2},\ D_{\max}=\mu+n\sigma\ (n=1,2,3) \tag{4}$$

The searching time of our normalization method increases linearly depending on the number of the points for the “Model” data. Initially, we use the skeleton with a lower resolution for some iteration in the ICP algorithm. And when it arrives at a convergence point, we apply all possible skeletons to the normalization process.

4 Global and Local Shape Comparisons

We estimate the shape difference by computing the distance for the sample meshes extracted from the skeleton points and the centers of sample meshes using the L_2 norm metric. Fig. 3 shows the results of comparison between the hippocampal structures of the normal control and a patient with epilepsy. Fig. 3a shows the result of global shape comparison. Fig. 3b and 3c show how to compare two hippocampal shapes based on the proposed Octree and skeletal scheme. It is possible to reduce the computation time in comparing two shapes by picking a certain skeletal point (Fig. 3b) or by localizing an Octree node (Fig. 3c) from the remaining parts. It is also possible to analyze the more detail region by expanding the resolution of the Octree, since it has a hierarchical structure. The result of shape comparison is displayed on the surface of the target object using color-coding.

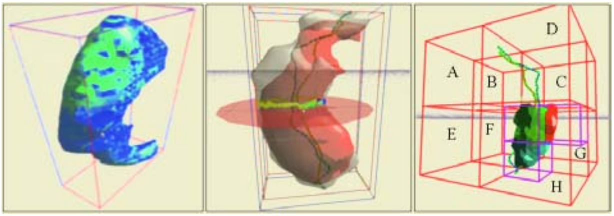


Fig. 3. Global and local shape comparison: (a) global comparison; (b) skeletal point picking based local comparison; (c) Octree-based hierarchical comparison

A great deal of effort is focused on the development of neural networks for various applications in medical areas [8]. In this paper, we have implemented a neural network classifier to discriminate between the normal controls and the epileptic patients. Fig. 4 is a system structure of the back-propagation neural network for a shape classification. We use skeletons and sample meshes to compute the input of a classifier. The input layer's values are composed of the distances from skeleton points to the sample meshes. The output layer has two kinds of values - 0 or 1 (where '0' means "epilepsy patients" and '1' means "normal controls").

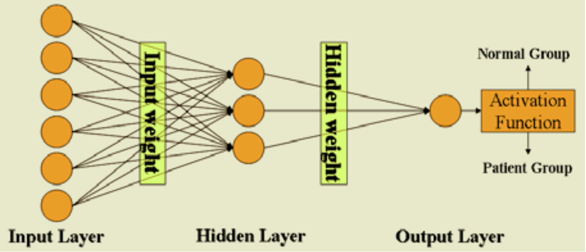
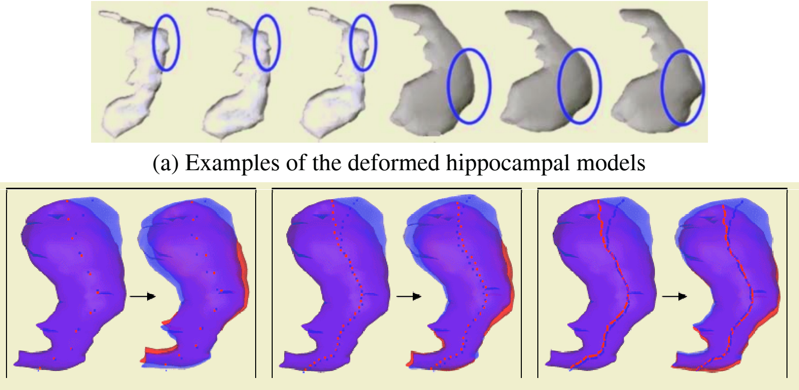


Fig. 4. Back-propagation neural network for the shape classification

5 Experimental Results

The present method was applied to analyze the hippocampal models reconstructed from the MR brain images. To evaluate the performance of the proposed normalization algorithm with the neural network based classifier, we reconstructed hippocampal models of a normal control and an epileptic patient using MRI and generated 80 deformed models. Fig. 5a shows some examples of the hippocampal models used in our experiments. The three left models are the epileptic cases and the three right models the normal cases. In each model, the blue circle indicates the deformed region of the hippocampus. Figs. 5b-5d show the results of ICP-based normalization as the number of skeletal points are increased. The fixed shape and the moveable shape are colored with red and blue, respectively. Within each image, the right party is the result of the normalization. Result indicates that the skeleton-based normalization using ICP is relatively stable over the resolution of skeleton. And we can reduce the computation time for the normalization by using the low-resolution skeleton as shown in Fig. 6, which shows the comparison result of the time efficiency in ICP-based normalization by varying the number of skeletal points.



(a) Examples of the deformed hippocampal models

(b) 11 skeletal points (c) 31 skeletal points (d) 101 skeletal points

Fig. 5. The results of normalization depending on the number of skeletal points

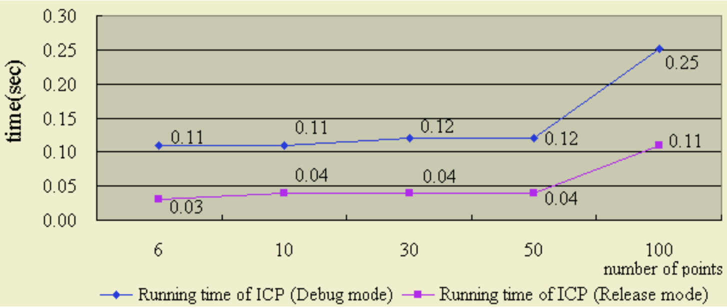


Fig. 6. The comparison of the ICP performance depending on the number of skeletal points

Table 1 gives the result of global shape differences between the normal left hippocampus (N_L) and three targets (T1, T2, and T3) according to L_2 norm and volume difference. As shown in Fig. 3c, we are able to evaluate the qualitative result of the shape difference at specific region and to control the hierarchical analysis using the Octree structure (i.e. upper-front-right, bottom-front-left, upper-back-left, and the bottom region, respectively). In Table 2, highlighted cells represent the deformed area of the hippocampus model and so we observe that the similarity error at deformed region is higher than at other regions. As shown in Tables 1 and 2, our method is able to discriminate the global shape difference and is able to distinguish a certain shape difference at a specific local region in a hierarchical fashion. To confirm the capacity of our classifier, we organized experimental data set by using the cross-validation technique. Table 3 shows the result of the classification based on the neural network.

Table 1. The result of the global shape analysis

	L_2 norm	Volume difference	Rank
N_L:T1	1.220	94.3%	1
N_L:T2	1.554	109.3%	2
N_L:T3	2.420	88.8%	3

Table 2. The result of local shape analysis based on the Octree structure

	A	B	C	D	E	F	G	H
P_L:T4	0.15	0.77	0.84	3.15	0.00	0.00	0.00	0.15
P_L:T5	1.20	0.00	0.00	0.00	3.12	2.00	1.00	1.44
N_R:T6	0.06	1.02	0.06	0.00	0.00	0.12	0.00	0.00
N_R:T7	0.00	0.00	0.00	0.00	1.54	1.31	1.313	1.54

Table 3. The result of the neural network based classification for the hippocampal models

Training Times	Output Error	Precision
720	0.014416	50%
10800	0.000339	98%
14400	0.000061	100%

6 Conclusions and Future Works

This paper presented an efficient shape representation for the analysis of hippocampal structure, where three different representations, i.e. meshes, voxels, and skeletons are combined in a hybrid fashion. As we integrate the hybrid representations into the Octree structure, we can efficiently estimate the global and local shape similarity of the hippocampus. In addition, the ICP-based normalization provided the fast and the accurate registration between two 3D models. Therefore, we could increase the speed of analysis without degrading accuracy using a level-of-detail approach. We also could discriminate between the normal controls and the epileptic patients using a neural network based classifier. In the near future, we aim to evaluate our method by collecting more hippocampal patient data. We are also considering to us Support Vector Machines for the classification task.

Acknowledgements

We would like to thank Prof. Seung-Bong Hong and Woo-Suk Tae in Samsung Medical Center for giving useful comments. This work was supported in part by the

Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2005-205-D00105). This work was supported in part by the Korean Ministry of Science and Technology under the NRL program.

References

1. Dean, D., Buckley, P., Bookstein, F., Kamath, J., Kwon, D., Friedman, L., Lys, C.: Three dimensional MR-based morphometric comparison of schizophrenic and normal cerebral ventricles. Vis. In Biom. Computing, Lecture Notes in Comp. Sc., 363-372, (1996)
2. Petrou, M. and Bosdogianni, P.: Image Processing: The Fundamentals, John Wiley, (1999)
3. Vranic, D.V. and Saupe, D.: 3D Model Retrieval, Proc. SCCG 2000, May 3-6, Budmerice, Slovakia, 89, (2000)
4. Paquet, E. and Rioux, M.: Nefertiti: a Query by Content System for Three-Dimensional Model and Image Databases Management, Image and Vision Computing, Vol. 17: 157, (1999)
5. William, E.L. and Harvy, E. C.: Marching cubes: A high resolution 3D surface construction algorithm. Computer Graphics, Vol. 21. No. 4:163-169, (1987)
6. Besl, P.J. and McKay, N.D.: A Method for Registration of 3D Shapes, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 14, No. 2: 239-256, (1992)
7. Zhang, Z.: Iterative point matching for registration of freeform curves and surfaces. International Journal of Computer Vision, Vol. 13, No. 2:119-152, (1994)
8. Sordo, M.: Introduction to Neural Networks in Healthcare, Open Clinical Document, October, (2002)

A Fast Image Retrieval Using the Unification Search Method of Binary Classification and Dimensionality Condensation of Feature Vectors

Jungwon Cho[1], Seungdo Jeong[2], and Byunguk Choi[3]

[1] Department of Computer Education, College of Education,
Cheju National University, 66 Jejudaehakno, Jeju-si, Jeju-do, 690-756 Korea
jwcho@cheju.ac.kr

[2] Multimedia Laboratory, Department of Electrical and Computer Engineering,
Hanyang University, 17 Haengdang-dong, Sungdong-gu, Seoul, 133-791 Korea
kain@mlab.hanyang.ac.kr

[3] Division of Information and Communications, Hanyang University,
17 Haengdang-dong, Sungdong-gu, Seoul, 133-791 Korea
buchoi@mlab.hanyang.ac.kr

Abstract. We present the two-stage content-based image retrieval as a new fast image retrieval approach using the unification search method of binary classification and dimensionality condensation of feature vectors. The method successfully reduces the overall retrieval time, while maintaining the same retrieval relevance as the conventional exhaustive search method. By the extensive computer simulations, we have observed that the method is more effective as user-specific threshold for the similarity score increase.

1 Introduction

The conventional exhaustive search methods in the content-based image retrieval (CBIR) compare the feature vector of the query with all feature vectors in the image feature database. Moreover, the CBIR system relies heavily on the similarity search over high dimensional spaces in large image database [1]. The larger the number of images is, and the higher the dimension of the feature spaces is, the greater the overall retrieval time of the retrieval systems becomes. The binary classification and the dimensionality condensation of feature vectors are essential for reducing the overall processing time for similarity computation. The binary classification is accomplished by grouping into the groups with similar image features, and the dimension condensation is accomplished by making the mapping from a high dimensional to a low dimensional feature space. There are two widely used approaches in the fast search method of the CBIR: the dimensionality reduction of the feature spaces [2–7] and the hierarchical clustering of similar images [8]. Although the dimensionality reduction methods such as the variable-subset selection [2], the multidimensional scaling [3, 4], the geometric hashing [5, 6], and the magnitude and shape approximations [7] are successfully

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 240–247, 2005.

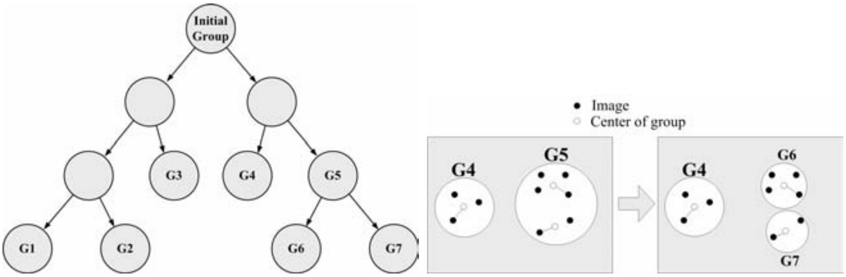
able to reduce the overall retrieval time by performing the search on a reduced dimensional feature space, they may alter the ranks of the retrieval results according to the change of the search space metric structure. The approach in [8] also shows the efficient overall retrieval time, but does not guarantee the same level of precision and recall ratios as the conventional exhaustive search case.

This paper presents the unification search method in order to reduce the overall retrieval time, while maintaining the retrieval relevance. The method uses both the binary classification and the dimensionality condensation. Given the image feature descriptor and similarity measure, the proposed method groups the similar images by making use of the binary classification based on the K-means algorithm and obtains the centers of all the groups. Also, the method condenses a high dimensional feature vector space into a low dimensional feature vector space by making use of Cauchy-Schwarz inequality. In order to confirm the original retrieval relevance, we suggest the two-stage search method, and compare the results with those of the conventional exhaustive image search in the database.

2 Building of Indexing Database

2.1 Binary Classification

The K-means algorithm classifies whole data into K groups [8]. However, it is almost impossible to find the best K in advance, therefore we need a adaptive method to find the best K dynamically. In this paper, we propose the binary classification method to find the best K dynamically. Fig. 1(a) shows the proposed method which is able to determine the number of groups dynamically by classifying groups which are not satisfied with the classification similarity threshold into the binary tree form. Also, Fig. 1(b) shows that G5 which is not satisfied with the classification similarity threshold is divided into G6 and G7 by the binary classification method.



(a) A form of binary tree (b) An example of classification

Fig. 1. Binary classification

Let μ_c be the predetermined user-specific classification similarity threshold. The initial group includes all images in the original image database and the group center is an average value of all feature vectors. If the maximum distance between the group center and image feature vectors is larger than μ_c, then the binary classification method is performed. Otherwise, the group is inserted into the representative vector database.

The binary classification method can be expressed as:

1. Binary classification.
   (a) Set ic (iteration count) to 1.
   (b) Choose randomly a set of two means $m_1(1), m_2(1)$ for binary classification.
   (c) For each feature vector d_i compute $\mu_{ik} = D(d_i, m_k(ic))$ for each $k = 1, 2$ and assign d_i to the group G_k with the nearest mean (minimum μ_{ik}).
   (d) Increment ic by 1.
   (e) Update the means to get a new set $m_1(ic), m_2(ic)$.
   (f) Repeat from c to e until $G_k(ic) = G_k(ic+1)$ for all k.
2. For each feature vector in G_k, d_{ki} compute $\mu_k = MAX(D(d_{ki}, m_k))$ for each $k = 1, 2$.
3. If μ_k is larger than μ_c, then push G_k onto the classification stack for binary classification; otherwise, insert G_k into the representative vector database.
4. Stop when the classification stack empty. (terminated condition).
5. Pop G_i, the top group from the classification stack.
6. Repeat from 1 to 5 for G_k.

2.2 Dimensionality Condensation

The Cauchy-Schwarz inequality is given by

$$\sqrt{\sum_j (q_j)^2 \sum_j (a_j)^2} \geq \sum_j q_j a_j \tag{1}$$

where q_j and a_j denote the j-th elements of the query feature vector and the image vector in the database, respectively, having arbitrary real values between 0 and 1. Note that the right-hand side (RHS) of Eq. (1) represents the inner product similarity metric. Since Eq. (1) always holds with all real numbers, if the inner product between the query feature vector and the image vector given in the RHS of Eq. (1) is greater than the predetermined user-specific threshold, say α, then the LHS of Eq. (1) is also greater than α.

For each of N images in the database, the dimensions of the condensed feature vector and the original feature vector are given by L and M, respectively. q_j $(1 \leq j \leq M)$ and c_j $(1 \leq j \leq L)$ denote the j-th elements of the query vector and its condensed vector, respectively. Also, $a_{i,j}$ $(1 \leq i \leq N,\ 1 \leq j \leq M)$ and $r_{i,j}$ $(1 \leq i \leq N,\ 1 \leq j \leq L)$ denote the j-th element of the i-th original image vector and its condensed vector, respectively. Note that the elements in the condensed domain are computed for $1 \leq j \leq L$ as

$$c_j = \sqrt{\sum_{k=(j-1)\lambda+1}^{(j-1)\lambda+\lambda} (q_k)^2},\ r_{i,j} = \sqrt{\sum_{k=(j-1)\lambda+1}^{(j-1)\lambda+\lambda} (a_{i,k})^2} \tag{2}$$

where λ represents the condensation ratio, defined as M/L [10].

3 Two-Stage Search Method

Fig. 2 illustrates a schematic diagram of the proposed CBIR system using the unification search method.

In the first stage search, the user-specific similarity threshold is readjusted by the similarity compensation, for reducing the overall retrieval time, the query is compared with the centers of all the groups in the representative vector database in Part A, and the candidate groups having larger similarity than the compensated threshold are selected. The query is then compared to the low dimensional condensed features of all the images in the candidate groups in Part B, and the candidate images having the large similarity to the query are selected. In the second stage search, for maintaining the retrieval relevance, the corresponding high dimensional original features with the candidate images from the first stage search are then compared to the query image.

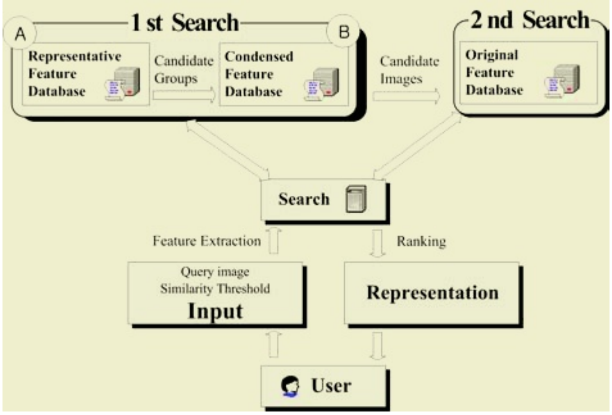


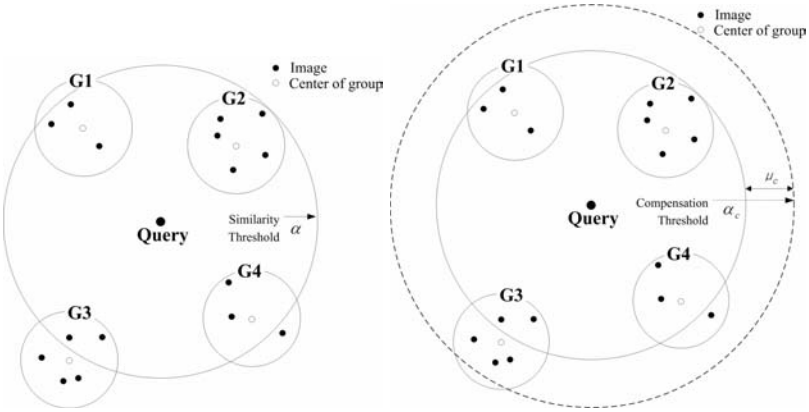
Fig. 2. Schematic diagram of the two-stage CBIR system

3.1 Similarity Compensation

Fig. 3 illustrates the compensation of the similarity threshold for maintaining the original retrieval relevance. In Fig. 3, (a) and (b) illustrate the states which the similarity compensation is performed and not performed, respectively. In Fig. 3(a), because the center of G3 is not satisfied with α, G3 is not selected as the candidate group. Though some images of G3 must be included in the final retrieval results because they are satisfied with α, they are excluded. To achieve the same relevance as the conventional exhaustive search method, α must be compensated shown as Fig. 3(b). Note that the compensated similarity threshold, say α_c, is computed under the Euclidean distance similarity metric as

$$\alpha_c = \alpha + \mu_c \tag{3}$$

Fig. 3(b) shows that G3 is determined as the candidate group when α_c is used. By the two-stage search method using α_c, we are successfully able to reduce the



(a) Before compensation - excluding G3

(b) After compensation - including G3

Fig. 3. Compensation of the similarity threshold

overall retrieval time and guarantee the same level of precision, recall and rank as the conventional exhaustive search method.

3.2 Search Using Condensed Feature Vectors

Let α be the predetermined user-specific threshold for the similarity score under the inner-product similarity metric, having the value between 0 and 1. Now, using Eq. (2) in Eq. (1), it is not difficult to show that for the i-th image

$$\sqrt{\sum_{j=1}^{M}(q_j)^2\sum_{j=1}^{M}(a_{i,j})^2} \geq \sum_{j=1}^{L} c_j r_{i,j} \geq \sum_{j=1}^{M} q_j a_{i,j} \geq \alpha. \tag{4}$$

For all i, in the first stage search, we select the candidate images by Part B based on the condition

$$\sum_{j=1}^{L} c_j r_{i,j} \geq \alpha. \tag{5}$$

In the second stage search, the final images are then obtained by using the corresponding candidate images for selected i according to the condition

$$\sum_{j=1}^{M} q_j a_{i,j} \geq \alpha. \tag{6}$$

From Eq. (4), since the condition in Eq. (5) is a necessary condition for the condition in Eq. (6), we are able to see that the candidate images obtained by Eq. (5) in the first stage for all i must contain the result images to be obtained by Eq. (6) for selected i. Note that the result images so obtained must be identical to the images obtained by the conventional exhaustive search method for all i.

Under the Euclidean distance metric, the conditions for selecting the candidate images by Part B of the first stage search and the final retrieved images by the second stage search will respectively be as follows:

$$\sqrt{\sum_{j=1}^{L} (c_j - r_{i,j})^2} \leq \alpha \tag{7}$$

$$\sqrt{\sum_{j=1}^{M} (q_j - a_{i,j})^2} \leq \alpha \tag{8}$$

Note that the conditions in Eq. (7) and Eq. (8) all satisfy the Cauchy-Schwarz inequality in Eq. (1).

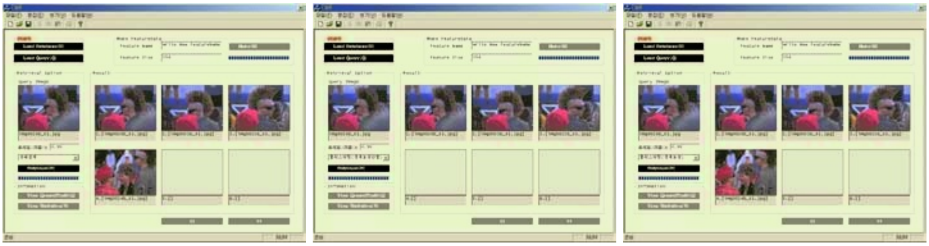
4 Experimental Results

For the experiment, we use the MPEG-7 color-structure descriptor [11]. The common color dataset (CCD) consisting of 5,466 images and a set of 50 common color queries (CCQ) are employed on a Pentium III-700 PC equipped with Windows 2000.

We want to verify experimentally that the proposed two-stage search method operates faster than the conventional exhaustive search method, while maintaining the same retrieval relevance. For this purpose, we only compare the performance of the proposed method with the performance of the conventional exhaustive search method.

Fig. 4 illustrates the comparison of the retrieval results from the search methods. Fig. 4(a) is the retrieval result from the conventional exhaustive search method, which is performed with all of images in the original image database. Both Fig. 4(b) and Fig. 4(c) are the retrieval results from the proposed two-stage search method. These results show before and after performing the similarity compensation, respectively. In Fig. 4(b), the recall ratio is decreased because the similarity compensation is not performed. In case of Fig. 4(c), however, we get the same result as Fig. 4(a) with performing the similarity compensation. Like the experiment shown as Fig. 4, we have confirmed empirically that our proposed two-stage search method yielded the same precision, the same recall ratio, and the same rank as the conventional exhaustive search method for all the 50 queries of CCQ. Hence, we only compare the overall retrieval time of the proposed method with that of the conventional exhaustive search method.

For the Euclidean distance metrics, the relative processing time is computed and tabulated in Table 1 for different choices of α. Here, by the relative processing time, we mean the ratio (in percentage) of the overall processing time for similarity computation between our proposed methods and the conventional exhaustive search method. In the table, the predetermined classification similarity threshold was selected to be $\mu_c = 0.9$, the condensation ratio was selected to be $\lambda = 8$ $(M = 256, L = 32)$.



(a) conventional exhaustive method (b) proposed - before compensation (c) proposed - after compensation

Fig. 4. The comparison of retrieval result

Table 1. The relative processing time for different values of α

α	Binary Classification	Dimensionality condensation	Unification method
0.7	30.64%	44.84%	21.58%
0.8	17.01%	25.46%	7.46%
0.9	10.37%	14.21%	1.92%

We observed that for larger α, the savings in overall processing time for similarity computation by our two-stage search method become larger comparing with the conventional exhaustive search method. This is because as α increases, the number of candidate images obtained from the first stage search decreases, and thus the processing time in the second stage search also decreases. It is important to notice that in order to reduce the overall processing time, the savings in processing time in the second stage search are more beneficial than those in the first stage search.

5 Conclusion

In this paper, we have proposed a new fast image retrieval using the two-stage search method based on the unification search method. The unification method uses both the binary classification and the dimensionality condensation of feature vectors. In order to maintain the same retrieval relevance as the conventional exhaustive search method, we have employed the similarity compensation and the two-stage search method. Extensive computer simulations have been performed in order to demonstrate the fast operation characteristics of our proposed method using the unification search method. We have confirmed that the method reduces the overall retrieval time quite successfully while maintaining the same retrieval relevance as the conventional exhaustive search method. We are now working on analytically finding the optimal value of condensation ratio and the similarity compensation.

Acknowledgement

This work was supported by HY-SDR Research Center at Hanyang University, Seoul, Korea, under the ITRC Program of IITA, Korea.

References

1. V. Castelli and L.D. Bergman: Image Databases. Jon Wiley & Sons, Inc. (2002)
2. B.V. Bonnlander, et. al.: Selecting input variables using mutual information and nonparametric density estimation. In Proc. of Int. Symp. on Artificial Neural Networks. (1994) 312–321
3. M. Beatty and B.S. Manjunath: Dimensionality reduction using multidimensional scaling for content-based retrieval. In Proc. of IEEE Int. Conf. on Image Processing. (1997) 835–838
4. C. Faloutsos and K.-I. Lin: FastMap: a fast algorithm for indexing, data-mining, and visualization of traditional and multimedia data sets. In Proc. of 1995 ACM SIGMOD Int. Conf. on Management of Data. (1995) 163–174
5. A. Califano and R. Mohan: Multidimensional Indexing for Recognizing Visual Shapes. IEEE trans. on Pattern Analysis & Machine Intelligence. Vol. 16, No. 4 (1994) 373–392
6. H. Wolfson and I. Rigoutsos: Geometric Hashing: An Overview. IEEE Computational Science & Engineering. Vol. 4 (1997) 10–21
7. U. Y. Ogras and H. Ferhatosmanoglu: Dimensionality Reduction Using Magnitude and Shape Approximations. In Proc. the Twelfth International Conference on Information and Knowledge Management. (2003) 99–107
8. S. Krishnamachari and M. Abdel-Mottaleb: Hierarchical clustering algorithms for fast image retrieval. IS&T/SPIE Conf. on Storage and Retrieval for Image and Video Databases VII. Vol. 3656 (1999) 427–435
9. L.G. Shapiro and G.C. Stockman: Computer Vision. Prentice Hall. (2001)
10. J. Cho, S. Jeong and B. Choi: A New Fast Image Retrieval Using the Condensed Two-Stage Search Method. IEICE Trans. on Communication. Vol. E86-B, No. 12 (2003) 3658–3661
11. B. Mamjunath, et. al.: Color and Texture Descriptors. IEEE Trans. on Circuits and Systems for Video Technology. Vol. 11, No. 6 (2001) 703–715

Semantic Supervised Clustering to Land Classification in Geo-Images

Miguel Torres, G. Guzman, Rolando Quintero,
Marco Moreno, and Serguei Levachkine

Geoprocessing Laboratory – Centre for Computing Research
National Polytechnic Institute, Mexico City, Mexico
{mtorres,jguzmanl,quintero,marcomoreno,sergei}@cic.ipn.mx
http://geo.cic.ipn.mx, http://geopro.cic.ipn.mx

Abstract. In this paper, we propose a semantic supervised clustering approach to classify lands in geo-images. We use the *Maximum Likelihood Method* to generate the clustering. In addition, we complement the analysis applying *spatial semantics* to improve the classification. The approach considers the *a priori* knowledge of the multispectral image to define the training sites (classes) related to the geographic environment. In this case the spatial semantics is defined by the spatial properties, functions and relations that involve the geo-image. By using these characteristics, it is possible to determine the training data sites with *a priori* knowledge. This method attempts to improve the supervised clustering, adding the intrinsic semantics of the geo-images to determine the training sites that involve the analysis with more precision.

1 Introduction

The integration of remote sensing and geographic information systems (GIS) in environmental applications has become increasingly common in recent years. Remotely sensed images are an important data source for environmental GIS-applications, and conversely GIS capabilities are being used to improve image analysis procedures. In fact, when image processing and GIS facilities are combined in an integrated system vector data, they can be used to assist in image classification and raster image statistics. In addition, vectors are used as criteria for spatial queries and analysis [1].

Frequently the need arises to analyze mixed spatial data. These data sets can consist of satellite spectral, topographic and other point form data, which are registered geometrically [2].

In this paper, we propose a semantic supervised clustering algorithm applied to LandSat TM images to classify the land, according to the *a priori* knowledge. The classification method is used to define the *training sites* considering the characteristics and intrinsic properties (spatial semantics). Our approach attempts to overcome some of the limitations associated with computational issues from previous supervised clustering methods.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 248–254, 2005.

Moreover, our approach is based on the *spatial semantics*[1], which is used to determine the behavior of a certain geographic environment by means of spatial properties, functions and relations [3].

The rest of the paper is organized as follows. In section 2 we describe the supervised clustering approach and the proposed algorithms to determine the training sites and the supervised clustering. Section 3 shows the results obtained by applying the approach to LandSat TM images. Our conclusions are outlined in section 4.

2 Supervised Clustering Approach

2.1 Supervised Clustering Method

The supervised clustering is the procedure used for quantitative analysis of remote sensing image data [4]. In this work, we have implemented the *Maximum Likelihood Classification* method. In common cases, the supervised clustering is defined through training sites, which are determined by pixels according to *a priori* semi-automatic selection.

We have considered five basic steps to generate a supervised clustering: (I) Determine the number and type of classes to use for the analysis, (II) Choose training regions (sites) for the classes, according to the *spatial semantics* to identify the spectral characteristics for each specific class, (III) Use these training regions to determine the parameters of the supervised clustering, (IV) Classify all the pixels from the geo-image, assigning them to one of the defined classes by the training regions, and (V) Summarize the results of the supervised clustering [5]. An important assumption in supervised clustering usually adopted in remote sensing is that each spectral class can be described by a probability distribution in a multispectral space. A multidimensional normal distribution is described as a function of a vector location in multispectral space by Eqn 1.

$$p(x) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right\}, \tag{1}$$

where:

N is the number of components of vector x.
x is a vector location in the N dimensional pixel space.
μ is the mean position of the spectral class.
$\sum$ is the covariance matrix of the distribution.

The multidimensional normal distribution is completely specified by its *mean vector* and its *covariance matrix* ($\sum$). In fact, if the mean vectors and covariance matrices are known by each spectral class, it is possible to compute the set of probabilities that describe the relative *likelihoods* of a pattern, at a particular location belonging to

[1] To define *spatial semantics* [2], we use the essential characteristics that involve the remote sensing imagery. Our definition is based on providing a set of rules that describe the geo-images. This set is composed of relations, properties and functions that define the behaviors of the raster image. This definition starts with a spatial description of the geographical objects that consists of topological, logical, geometrical and statistical properties, which define conceptually spatial semantic rules that involve the geo-image

each of these classes [6]. This can be considered as a member of the class, which indicates the highest probability. Therefore, μ and ∑ are known by every spectral class in an image, also every pixel can be examined and labeled according to the probabilities computed for the particular location of every pixel. Before the classification, μ and ∑ are estimated for each class from a representative set of pixels, commonly called *training sites*.

2.2 Detection of Semantic Training Sites Algorithm

The main issue of our research is to obtain the training sites (classes) to determine the areas that involve the classification. In addition, we attempt to define these areas by means of the spatial semantics[2] of the geo-images to improve the classification method. The semantic clustering algorithm is described as follows:

[Step 1]. Let O be the sampling set of objects. See Eqn. 2.

$$O = \{o_i \mid o_i \in Z^2\}, \tag{2}$$

Hence each vector o_i is a 3-tuple, which contains the RGB values[3] at the position (x, y) of the object selected randomly in the geo-image. According to the tests, the number of *seeds (i)* that provides good results is shown in Eqn. 3.

$$k_1(M \cdot N) \leq i \leq k_2(M \cdot N), \tag{3}$$

where:

M is the number of rows in the geo-image.
N is the number of columns in the geo-image.
K_1 = 0.01is the proposed value of seeds for every class (ω).
K_2 = 0.03 is the proposed value of seeds for every class (ω).

[Step 2]. Transform each vector o_i from spatial domain Z^2 to the semantic domain S by means of Eqn. 4.

$$S_O = \{s_i \mid s(o_i) \neg o_i \in O\} \tag{4}$$

Therefore, *s(o)* is the semantic characteristic vector, which is obtained by applying Eqn 5.

$$s = < s_1, s_2, s_3, s_4, s_5, s_6 >, \tag{5}$$

where: s_1, s_2, s_3 are the mean of the original seed and their 8-neighbors[4] in the spatial domain Z^2, according to the RGB spectral bands respectively. Additionally s_4, s_5, s_6 are the values of the standard deviation[5] of the original seed and their 8-neighbors in the same domain.

[2] In the semantic clustering algorithm, we consider the "essential" properties that describe the semantics of the geo-images. The essential features for this case of study are the topological properties, which are considered the *a priori* knowledge

[3] In other types of images, we can use other spectral components

[4] In this view, it is important to define the topological property of the geo-image, because this property is considered the semantics of the raster data

[5] Standard deviation represents the variability of the reflectance in every selected spectral band

[Step 3]. Obtain the dissimilitude coefficient (DSC) *d(s$_i$, s$_j$)* for *s$_i$, s$_j$* that belongs to *S$_o$*, which is defined in Eqn. 6.

$$d(s_x, s_y) = \sqrt{\left(s_{x1}^{\ 2} - s_{y1}^{\ 2}\right) + \left(s_{x2}^{\ 2} - s_{y2}^{\ 2}\right) + \ldots + \left(s_{x6}^{\ 2} - s_{y6}^{\ 2}\right)} \tag{6}$$

[Step 4]. Let $D = \{d(s_i, s_j) \mid s_i, s_j \in S_o \forall i \neq j\}$ be the set of DSCs, and $d_M = Max(D)$ and $d_m = Min(D)$. The threshold of maxim similarity *U* is obtained by applying Eqn. 7.

$$U = \lambda(d_M - d_m), \tag{7}$$

where:

λ is the discrimination coefficient[6].

[Step 5]. Let *s$_i$*, *s$_j$* and *s$_k$* be vectors that belong to *S$_o$*, $s_k = f(s_i, s_j)$ in which *S$_o$* is the merging process of *s$_i$* and *s$_j$*; if Eqn. 8 is accomplished.

$$s_k = \left\langle \frac{s_{i1} + s_{j1}}{2}, \frac{s_{i2} + s_{j2}}{2}, \ldots, \frac{s_{i6} + s_{j6}}{2} \right\rangle \tag{8}$$

Let *g:S$_o$* → *S$_o$* be the minimal dissimilitude function (MDS) in which *g(s$_i$)* is the most similar vector to *s$_i$*. Then an evolution of the set *S$_o$* is defined by Eqn. 9.

$$\begin{aligned} S_o^{'} &= \{s_i \mid g(g(s_i)) \neq s_i \forall s_i \in S_o\} \cup \\ &\{s_f = f(s_i, s_j) \mid g(s_i) = s_j, g(s_j) = s_i \forall s_i, s_j \in S_o\} \end{aligned} \tag{9}$$

If the dissimilitude distance of a vector *s$_i$* is greater than the threshold *U*, regarding the rest of the vectors, then *s$_i$* goes to the next generation of vector. Otherwise, vector *s$_i$* must be merged with other vectors, whose dissimilitude distance is less than the threshold *U*.

[Step 6]. If $card(S'_o) < M$, then repeat from step 4, with $\lambda = \lambda / 2$.

[Step 7]. Determine the proportional ratio of the classes in the semantic domain *S*, which is given by Eqn. 10.

$$P_r = \frac{d_m}{d_M}, \tag{10}$$

While the proportional ratio of the classes is closer to 1, the partition of the semantic domain will be more proportional; that means that the dissimilitude distances between vectors are closer.

[Step 8]. Since the process is iterative, it should be repeated with the new generation of vectors, which have been obtained; that is, *S$_o$ = S$_o$'*

[Step 9]. Repeat the process from step 3, until *card(S$_o$) = M*.

2.3 Semantic Supervised Clustering (SSC) Algorithm

The result of the last stage (section 2.2) is the set of semantic characteristic vectors ω_i of the classification. From the Ω set, it is necessary to compute the mean and covari-

[6] The initial value for λ is 0.5. This value provides good results according to the tests

ance matrix for every $\omega_i \in \Omega$. By attempting to determine the class or category for every pixel at a location *x*, it is necessary that every pixel contain a conditional probability, denoted by Eqn. 11:

$$p(\omega_i \mid x), i = 1,...,M \tag{11}$$

In [4] we describe the supervised clustering method for more details. Therefore, Bayes theorem provides potential means of converting knowledge of predictive correlations. The constraint (Eqn. 12) is used in the classification algorithm, since the $p(x_i)$ are known by training data, and it is conceivable that the $p(\omega_i)$ can be also known or estimated from the knowledge analysis of the geo-image.

$$x \in \omega_i \text{ if } p(x \mid \omega_i)p(\omega_i) > p(x \mid \omega_i)p(\omega_j) \text{for all } j \neq i \tag{12}$$

In this analysis, we assume that the classes have multidimensional normal distributions and each pixel can assign a probability of being a member of each class. After computing the probabilities of a pixel being in each of the available classes, we assign the class with the highest probability. The algorithm consists of the following steps:

[Step 1]. Determine the number of classes ω_i by means of the semantic training sites algorithm (section 2.2).

[Step 2]. Compute the maximum likelihood distribution and the covariance matrix for each generated class ω_i.

[Step 3]. For each image pixel, determine its semantic vector, applying Eqn. 5.

[Step 4]. Compute the probability of vector *s* to know if it belongs to each class ω_i.

[Step 5]. Obtain the coordinates (*x, y*) of the pixel, if the constraint (see Eqn. 13) is accomplished, then the pixel belongs to the class ω_i.

$$p(x \mid \omega_i)p(\omega_i) > p(x \mid \omega_j)p(\omega_j) \text{for all } j \neq i \tag{13}$$

[Step 6]. Repeat from step 3, until all pixels in the geo-image can be classified.

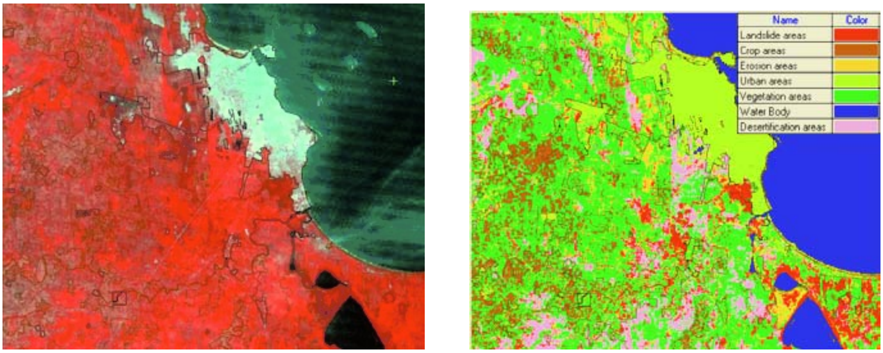
3 Results

By using this approach, we can perform a semantic supervised clustering in LandSat TM images. The algorithm has been implemented in C++ Builder. The segment of Veracruz State LandSat TM image is shown in Fig. 1. In Fig. 1a, we appreciate the LandSat TM image, which is composed of three spectral bands (4 3 2). In addition, we have overlapped vector data to identify the urban and land areas. Fig. 1b depicts the results of the semantic supervised clustering.

We have considered that to obtain a good estimation of class statistics, it is necessary to choose several trainings fields for the one cover type, located in different regions of the image. The three band signatures for the classes are obtained from the training fields, which are given in Table 1.

4 Conclusions

In the present work the *semantic supervised clustering* approach for Landsat TM (7 bands, resolution 25 meters per pixel) images is proposed. In addition, we propose to use *spatial semantics* to improve the classification by means of *a priori* knowledge, related to the selection criteria of the training sites (topological properties).



(a) (b)

Fig. 1. (a) Landsat image of Veracruz basin[7]. (b) SSC of Landsat image

Table 1. Class signatures generated from the training areas[8] in Fig. 1b. Numbers are in the scale of 0 to 255 (8 bits)

Class	Mean vector	Standard deviation	Covariance matrix		
Water body	39.16	1.35	1.82	0.00	0.00
	34.30	1.74	1.83	3.03	0.00
	21.38	2.43	2.24	3.40	5.91
Urban areas	52.27	5.11	26.15	0.00	0.00
	58.40	7.61	37.77	58.02	0.00
	58.50	7.31	25.61	37.84	53.56
Vegetation areas	44.83	1.30	1.69	0.00	0.00
	49.26	2.32	2.42	5.38	0.00
	62.79	5.36	4.65	8.36	28.76
Erosion areas	41.66	0.85	0.72	0.00	0.00
	42.81	1.28	0.50	1.66	0.00
	54.45	4.84	1.73	0.38	23.46

By applying this algorithm, it is essential to consider *a priori knowledge* involved in the training sites. We assume that the set of properties provides a starting point for the clustering process. The algorithm aims to preserve spectral properties and relations between information and spectral classes. The method is used to make a quantitative analysis of sensing spatial data.

Moreover, the semantic clustering algorithm consists of carrying out the problem from the spatial domain to the semantic domain. In this context, the most similar semantic classes will be merged according to their properties. This process is performed by means of computing the dissimilitude coefficient (DSC) and the merging of pairs of classes. With this approach, it is possible to reduce the number of classes to the desired number.

The semantic supervised clustering approach is essential for the decision making process. Additionally, it is used to improve the spatial analysis in different geographical environments. The approach allows us to make more accurate the spatial analysis

[7] Basin is the entire geographical area drained by a river and its tributaries

[8] Due to space limitations, we only present four generated classes in Table 1

with data fusion (vector and raster). This algorithm can be incorporated to detect flooding and landslide areas.

On the other hand, this approach can be used to know other useful spatial and attributive properties, which define the *spatial semantics* of geo-images.

Our future works are related to compare this method with others supervised clustering to evaluate the performance and the results of the raster data classification. In addition, we are looking for defining an *automatic* methodology to classify geo-images by means of *semantic unsupervised clustering*. For this purpose, it is necessary to use the *spatial semantics* to know the properties and behavior of raster data.

Acknowledgments

The author of this paper wishes to thank the CIC, CGPI, IPN and CONACYT for their support.

References

1. Unsalan, C.,Boyer, K.L.: Classifying land development in high-resolution Satellite imagery using hybrid structural-multispectral features. IEEE Transactions on GeoScience and Remote Sensing, Vol. 42, No. 12, (2004), pp. 2840-2850.
2. Torres, M., Levachkine S.: Generating spatial ontologies based on spatial semantics, in: Levachkine S., Serra J., Egenhofer M. (Eds.), Research on Computing Science, Semantic Processing of Spatial Data, Vol. 4, (2003), pp. 169-178.
3. Morgan, J.T., Ham, J., Crawford, M.M., Henneguelle, A., Ghosh, J.: Adaptative feature spaces for land cover classification with limited ground truth data. International Journal of Pattern Recognition and Artificial Intelligence, World Scientific Publishing Company, Vol. 18, No. 5, (2004), pp. 777-799.
4. Torres M., Moreno M., Quintero R., Guzmán G.: Applying Supervised Clustering to Landsat MSS Images into GIS-Application, Advances in: Artificial Intelligence, Computing Science and Computer Engineering, Research on Computing Science, Vol. 10, (2004), pp. 167-176.
5. Chung, K.F., Wang, S.T.: Note on the relationship between probabilistic and fuzzy clustering. Soft Computing, Lecture Notes in Computer Science, Springer-Verlag, Berlin Heidelberg, Vol. 8, No. 7, (2003), pp. 523-526.
6. Bandyopadhyay, S., Maulik, U., Pakhira, M.K.: Clustering using simulated annealing with probabilistic redistribution. International Journal of Pattern Recognition and Artificial Intelligence, World Scientific Publishing Company, Vol. 15, No. 2, (2001), 269-285.

Towards an Intelligent Web Service for Ontology-Based Query-Answering Dialogues

In-Cheol Kim

Department of Computer Science, Kyonggi University
Suwon-si, Kyonggi-do, 442-760, South Korea
kic@kyonggi.ac.kr

Abstract. In this paper, we present the design of an intelligent web service for ontology-based query-answering dialogues. This web service is realized through OWL-QL message exchanges between the querying agent and the answering agent. The OWL-QL is a formal language and protocol for query-answering dialogues among Semantic Web agents using knowledge represented in the standard Ontology Web Language(OWL). In OWL-QL settings, the answering agent uses automated reasoning methods to derive answers to queries. We explain the organizational details of each agent and then discuss the usefulness of the graphical OWL-QL query composer included in the querying agent.

1 Introduction

The semantic web is expected to include many kinds of query-answering services with access to many types of information represented in many formats [4]. Traditional database query languages like SQL and languages for retrieving information from the Web (e.g., XQuery and RQL[5]) are not suitable for supporting such heterogeneity, ranging from simple services that provide retrieval-based functionality to complex services that provide sophisticated automated reasoning functionality. OWL-QL [10] supports query-answering dialogues in which the answering agent may use automated reasoning methods to derive answers to queries, as well as scenarios in which the knowledge to be used in answering a query may be in multiple knowledge bases on the semantic web. Currently web services are emerging as a new breed of Web application [8]. They are self-contained, self-describing, modular applications that can be published, located, and invoked across the Web. Where the current web enables users to connect to applications, the web services architecture enables applications to connect to other applications [1]. In spite of providing high interoperability, however, current web service technology around SOAP, WSDL, and UDDI cannot make use of its full potential due to the lack of semantics. Semantic web service technology around OWL-S [3] and WSMO is being considered as a solution to address this limitation. In this paper, we present the design of an intelligent web service for ontology-based query-answering dialogues. This web service is realized through OWL-QL message exchanges between the querying agent and the answering agent. We explain the organizational details of each agent and then discuss the usefulness of the graphical OWL-QL query composer included in the querying agent.

2 OWL-QL Query Language

At present, the most important ontology languages are XML, XML Schema, RDF, RDF Schema, and OWL. Among them, RDF is a data model for objects ("resources")

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 255–260, 2005.

and relations between them, provides a simple semantics for this data model, and these data models can be represented in a XML syntax. RDF Schema is a vocabulary for describing properties and classes of RDF resources, with a semantics for generalization-hierarchies of such properties and classes. OWL adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties, characteristics of properties (e.g. symmetry), and enumerated classes [5].

There exist some query languages for XML documents, such as XPath, XQuery, and XQL. And there are also several query languages for RDF documents, such as RQL, RDQL, and Squish. OWL, however, is located at an upper level of abstraction than both XML and RDF. This fact would lead to complications if we were querying OWL documents with an XML or RDF-based language. So we need a new query language to write queries at the level of OWL. An appropriate query language must understand not only the syntax but also the semantics of OWL vocabulary. OWL-QL [10] is the proposed query language for OWL ontologies. OWL-QL follows a simple request-response model. The querying agent, or client, issues a query to the answering agent, or server. Suppose a simple OWL-QL message exchange:

If C1 is a Seafood Course and W1 is a drink of C1, what color is W1?
P: (type C1 Seafood-Course) (drink C1 W1)
Q: (has-color W1 ?x)
must-bind ?x
A: White

An OWL-QL query consists of premise, query pattern, variable bindings, and knowledge base(KB). The example query looks like the following:

```
<owl-ql:query xmlns:owl-ql="http://www.w3.org/2003/10/owl-ql-syntax#"
                  … xmlns:owl="http://www.w3.org/2002/07/owl#">
<owl-ql:premise>
  <rdf:RDF>
   <rdf:Description rdf:about="#C1">
    <rdf:type rdf:resource="#Seafood-Course"/>
    <drink rdf:resource="#W1"/>
   </rdf:Description>
  </rdf:RDF>
 </owl-ql:premise>
<owl-ql:queryPattern>
  <rdf:RDF>
   <rdf:Description rdf:about="#W1">
    <has-color rdf:resource="http://www.w3.org/2003/10/owl-ql-variables#x"/>
   </rdf:Description>
  </rdf:RDF>
 </owl-ql:queryPattern>
<owl-ql:mustBindVars>
  <var:x/>
 </owl-ql:mustBindVars>
<owl-ql:answerKBPattern>
  <owl-ql:kbRef rdf:resource="http://ontolingua.stanford.edu/wines.owl"/>
 </owl-ql:answerKBPattern>
</owl-ql:query>
```

An OWL-QL query performs some processing on the server, and produces a set of answer bundles. A server may divide up its answers into bundles separated by continuation tokens. The example answer looks like the following:

```
<owl-ql:answerBundle xmlns:owl-ql=http://www.w3.org/2003/10/owl-ql-syntax#
                              xmlns:var="http://www.w3.org/2003/10/owl-ql-variables#">
<owl-ql:answer>
  <owl-ql:binding-set>
   <var:x rdf:resource="#White"/>
  </owl-ql:binding-set>
  <owl-ql:answerPatternInstance>
   <rdf:RDF>
    <rdf:Description rdf:about="#W1">
     <has-color rdf:resource="#White"/>
    </rdf:Description>
   </rdf:RDF>
  </owl-ql:answerPatternInstance>
</owl-ql:answer>
</owl-ql:answerBundle>
```

3 Query-Answering Web Service

In view of Service Oriented Computing (SOC) [8], the OWL-QL answering agent can be considered as providing a certain kind of services. Based upon domain--specific knowledge bases (e.g. ontologies) and a powerful inference engine, it provides answering services in response to querying requests from distant querying agents. Wrapping the capability of the OWL-QL answering agent into a web service can increase its accessibility, reusability, and interoperability with other clients. Generally the web services architecture includes three different roles (e.g. service provider, service registry, and service requester) and the interactions between them (e.g. publish, find, and bind).

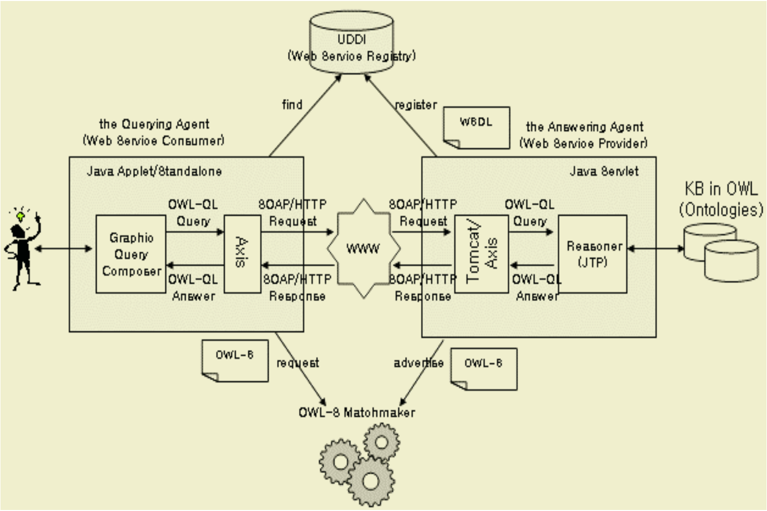


Fig. 1. Architecture of the Query-Answering Web Service

The web services architecture is based upon a set of XML-based open standards such as SOAP, WSDL, and UDDI. SOAP is W3C's recommended XML-data transport protocol, used for data exchange over web-based communication protocols such as HTTP. WSDL is the W3C recommended language for describing the service interface. UDDI defines a registry service for web services. Web services may be registered with a UDDI registry, which can subsequently be queried by other users and services.

The overall architecture of the OWL-QL query-answering web service is illustrated in Fig.1. The service architecture is comprised mainly of three components: the querying agent as a service consumer, the answering agent as a service provider, and an OWL-S matchmaker together with a UDDI registry as a service broker. The answering agent in the server-side, in turn, is composed of domain-specific knowledge bases (e.g. ontologies in OWL), a logic-based reasoner (e.g. JTP), an application server (e.g. Apache Tomcat), and a SOAP message engine (e.g. Apache Axis [2]). Here the reasoner, JTP[9], is adopted as OWL-QL query processor, which generates proper answers corresponding to the received queries. The current implementation of JTP supports OWL reasoning capabilities as well as RDF/RDF-S reasoning capabilities.

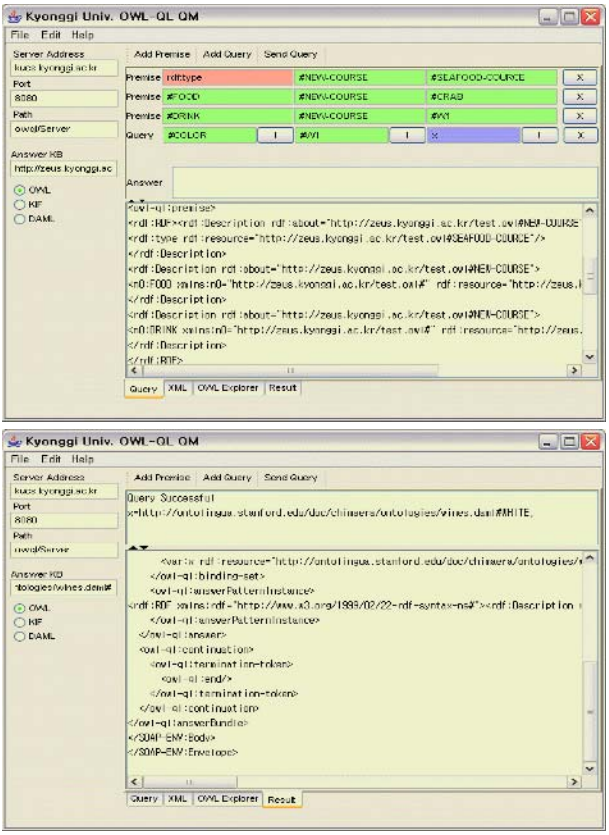
For example, it can imply the fact (owl:disjointWith Plant Mammal) from (owl:disjointWith Plant Animal) and (rdfs:subClassOf Mammal Animal) by enforcing the inheritance of owl:disjointWith constraints. The SOAP engine, Axis, plays key role in decoding the incoming SOAP request messages and encoding the outgoing SOAP response messages. On the other hand, the querying agent in the client-side consists of a query composer and a SOAP message engine(e.g. Apache Axis). The query composer includes a graphical query composer and a query converter. Unlike other ontology query languages such as RQL and RDQL, OWL-QL is a formal query language with an XML-based syntax. So, it is hard for humans to compose a query by manipulating it directly. In order to overcome this difficulty, our OWL-QL querying agent provides the user with a graphical query composer. Through a user-friendly interface, the user can compose his/her queries with ease. The query converter takes a query as input from the graphical query composer and translates it into the XML-based OWL-QL query. In this case, the SOAP engine, Axis, encodes the outgoing SOAP request messages and decodes the incoming SOAP response messages.

In addition to the WSDL description, our OWL-QL answering service also provides a semantic description represented in OWL-S. Basically, the OWL-S description defines three elements: a service profile to describe the functionality of the service, a service model to model the structure of the service, and a service grounding to map the abstract interface to a concrete binding information. By using the OWL-S descriptions, the matchmaker can performs a flexible capability-based service matching, but not just a simple keyword matching. Our OWL-QL answering agent advertises its service in an OWL-S description, and a service requester also queries for the service with an OWL-S description expressing its requirements. Then a service matchmaker (e.g. CMU's OWL-S4UDDI) may find matches between the service requirements and the advertised service according to their descriptions. Generally a matching algorithm proceeds in three stages: (a) the matching of service inputs, (b) the matching of service outputs, and (c) the matching of the service profile classification itself [6], [7]. Additionally, our OWL-QL answering agent provides a description of its service to a UDDI registry (e.g. IBM, HP, and XMethods UDDIs). This description includes a profile on the provider (e.g. university name and address); a profile about the service itself (e.g. name, category); and the URL of its WSDL description. A potential service requester may find our OWL-QL answering service through browsing and searching on the UDDI registry.

While the answering agent was implemented as a Java Servlet, the querying agent was implemented both as a Java Applet and as a Java standalone application. Fig.2 shows a screenshot of the querying agent in execution. On the left, a query is being composed within the graphical query composer. On the right, the returned answer is displayed in XML-based syntax.

4 Conclusions

In this paper, we presented the design of an intelligent web service for ontology-based query-answering dialogues. In our web service architecture, the querying agent communicates with the answering agent in OWL-QL. Our attempt to wrap the capability of the OWL-QL answering agent into a web service can greatly increase its accessibility, reusability, and interoperability with other clients.



Server Address: kucs.kyonggi.ac.kr
Port: 8080
Path: owl/Server
Answer KB: http://zeus.kyonggi.ac
● OWL
○ KIF
○ DAML

Premise	rdf:type	#NEW-COURSE	#SEAFOOD-COURCE	X
Premise	#FOOD	#NEW-COURSE	#CRAB	X
Premise	#DRINK	#NEW-COURSE	#W1	X
Query	#COLOR	#W1	x	X

Answer

```
<owl-ql:premise>
<rdf:RDF><rdf:Description rdf:about="http://zeus.kyonggi.ac.kr/test.owl#NEW-COURSE"
<rdf:type rdf:resource="http://zeus.kyonggi.ac.kr/test.owl#SEAFOOD-COURCE"/>
</rdf:Description>
<rdf:Description rdf:about="http://zeus.kyonggi.ac.kr/test.owl#NEW-COURSE">
<n0:FOOD xmlns:n0="http://zeus.kyonggi.ac.kr/test.owl#" rdf:resource="http://zeus.k
</rdf:Description>
<rdf:Description rdf:about="http://zeus.kyonggi.ac.kr/test.owl#NEW-COURSE">
<n0:DRINK xmlns:n0="http://zeus.kyonggi.ac.kr/test.owl#" rdf:resource="http://zeus.
</rdf:Description>
</rdf:RDF>
```

Server Address: kucs.kyonggi.ac.kr
Port: 8080
Path: owl/Server
Answer KB: ntologies/wines.daml#
● OWL
○ KIF
○ DAML

```
Query Successful
x=http://ontolingua.stanford.edu/doc/chimaera/ontologies/wines.daml#WHITE,
```

```
        <var:x rdf:resource="http://ontolingua.stanford.edu/doc/chimaera/ontologies/
      </owl-ql:binding-set>
      <owl-ql:answerPatternInstance>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"><rdf:Description
      </owl-ql:answerPatternInstance>
    </owl-ql:answer>
    <owl-ql:continuation>
      <owl-ql:termination-token>
        <owl-ql:end/>
      </owl-ql:termination-token>
    </owl-ql:continuation>
</owl-ql:answerBundle>
</SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

Fig. 2. A Screenshot of the Querying Agent: (Above) Query and (Below) Answer

Acknowledgements

This study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea. (No. 0412-MI02-0404-0002)

References

1. Booth D. et al.: Web Services Architecture, W3C Working Draft 8 August 2003, http://www.w3c.org/TR/2003/WD-ws-arch-20030808/ (2003)
2. Dave C., Tyler J.: Java Web Services, O'reilly (2002)
3. David M., Mark B., Ora L., Massimo P., Terry P., Sheila M.: Describing Web Services using OWL-S and WSDL, http://www.daml.org/services/owl-s/1.1/owl-s-wsdl.html, Nov. (2004)
4. Fensel D., Hendler J., Lieberman H., Wahlster W.: Spinning the Semantic Web, MIT Press (2003)
5. Grigoris A., Frank H.: A Semantic Web Primer, MIT Press. (2004)
6. Paolucci M. et al.: Semantic Matching of Web Services Capabilities, Proceedings of ISWC-2002, Lecture Notes in AI Volume 2342. (2002)
7. Paolucci M. et al.: Delivering Semantic Web Services, Proceedings of WWW-2003 (2003)
8. Munindar P. S., Michael N. H.: Service-Oriented Computing, Wiley (2005)
9. Richard F., Jenkins J., Frank G.: JTP: A System Architecture and Component Library for Hybrid Reasoning, KSL Technical Report 03-01. (2003)
10. Richard F., Pat H., Ian H.: OWL-QL: A Language for Deductive Query Answering on the Semantic Web, KSL Technical Report 03-14. (2003)

Using the Geographic Distance for Selecting the Nearest Agent in Intermediary-Based Access to Internet Resources

Leszek Borzemski and Ziemowit Nowak

Wroclaw University of Technology, Institute of Information Science and Engineering,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
{leszek.borzemski,ziemowit.nowak}@pwr.wroc.pl

Abstract. Intermediary agents extend the functionality and performance without violating the principles employed in the design of the Web. Such agents intervene in the client–server interactions shaping the response from the web server before it reaches the client. In this paper, we study and characterize the problem of choosing the intermediary agent nearest to the web client. Agent-client proximity is considered in the context of low latency and high bandwidth. Each intermediary agent can support every client but the nearest agent is the best choice. We performed extensive experiments on the Internet. We show that the geographic distance for selecting the nearest intermediary agent is accurate and effective.

1 Introduction

The Word Wide Web is based on the client-server model. The client sends a request and the Web server returns the response to the client. Usually, the server response does not depend on the parameters of the client. But along with development of the Internet technologies the responses are more often dependent on parameters of the client. One of such parameters is the client location, in particular its geographical location. In such case the server after receiving the request from the client determines the client location and sends the response which is dependent on client location.

In some cases the server is not able to generate the final response itself and needs some agents to prepare the final response. Such agents can be developed as intermediary systems that mediate the interaction between clients and servers of the World Wide Web [4]. They intervene in the client–server interactions. The purpose of intermediaries is to extend the functionality and performance without violating the principles employed in the design of the Web. The intermediaries are developed based on their functionality and focus. Nowadays we can use such intermediaries as message relays, caches, proxies, content mirrors, load balancers, protocol gateways, content filters and virus scanners. A typical operation involves the modification of original server responses and creation of the final response content which is returned to the client. Intermediaries can shape the server response in different ways taking into account the needs of local user community which is supported by that agent.

Most intermediary systems are “traditional” software in that they are not “smart” as intelligent ones. Intermediary intelligent system is characterized by its autonomy, mobility, adaptability, collaborative behaviour, reactivity, personality, operational continuity and knowledge-level communication ability. Intelligent agent has direct

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 261–267, 2005.

links with the problem environment. It is situated in user environment and is capable of sensing and affecting the state of the environment. It can make appropriate decisions on behalf of local user community when the state is changing.

Let us consider here the set of intermediary agents of the same functionality. Each intermediary agent can support every client but the nearest agent is the best choice. When using such intermediaries it is required to access the intermediary agent which is the nearest neighbour to the client in the context of low latency and high bandwidth. The implementation of such intermediary agents in many places of Internet might improve performance in access to WWW resources. Such intelligent agents can permanently monitor the network and decide which server is to be used at specific time moment.

Four typical Internet distance metrics can be used in the evaluation of the agent-to-client proximity: IP path length, time delay, AS path length and geographic distance. In this paper we show how the geographic distance can be usefully applied for selecting the best agent in such intermediary-based access to Internet resources.

The rest of this paper is organized as follows. Section 2 discusses related work in the Internet distance question. Section 3 introduces the centralized and distributed server shaping approaches. Section 4 presents the geographic distance calculation. Section 5 shows evaluation of the approach. Section 6 includes final remarks.

2 Related Work

The distance is most often expressed on the Internet by means of the measured delay (called RTT) in packet transmission from the source host to the target host, and back again. Indeed, in case of computer networks where the transmission is realized be passing the package through different transmission media and network devices the usage of such distance measure is a good choice. The time delay is being exploited on the Internet intensively as the measure of the distance. Thanks to well known ping service the measurement of the distance using this measure does not cause any problem when we are able to use ping on the source host.

Nevertheless if the distance information between any two arbitrary hosts in Internet is needed, the matter is growing complicated. One method to obtain distance information between any pair of hosts is shown in [5]. A global measurement infrastructure for Internet host distance estimation and distribution called Internet Distance Map Service (IDMaps) is proposed. Another proposal called GNP (Global Network Positioning) [10] introduces the model of the Internet as a geometric space and characterizes the position of a host by a point in this space. Knowing the co-ordinates of two hosts it is possible to calculate the distance between them. This approach was shown to be more accurate than IDMaps. However we cannot hope for both such services in the Internet. Moreover both services require a special global infrastructure. King is a service which is independent of the special infrastructure on the Internet [6]. This service can also estimate the latency between arbitrary Internet end hosts. The idea assumes that the authoritative DNS servers are located in most cases in the proximity of hosts which names are resolved. The authors state that effectiveness of King is exceeding both previous approaches.

Other Internet distance metrics can also be considered: IP path length; autonomous system (AS) path length and geographic distance. These metrics have been studied in comparison to time delay metric [8]. It was shown that metrics based on time delays yielded the best results (correct estimation was in 90% of trials). The geographic distance was the second best metric. It achieved a 75% success rate of trials. The results obtained for the geographical distance are important because when knowing geographical co-ordinates of hosts we can estimate Internet distance between these hosts in a real-time and what is very important without the need to measure the time delay.

The geographical distance on Internet was extensively studied in [7, 11]. Paper [9] formulates the tip that the median of time delays is a function of the geographical distance. But a problem is left how to determine the geographical localization of hosts on the Internet. The direct (using the measurements) method for determining the geographical location is mainly based on the Global Positioning System (GPS). However it has the basic shortcoming: in the location to be determined a special device has to be installed with GPS functionality. It would be an ideal method installing the GPS device in every host in the Internet. But this is not easy for implementation, therefore we look for other non-direct methods.

3 Centralized and Distributed Server Response Shaping

The intervention in the server response can be done in centralized or distributed manner. The centralized method was illustrated in Fig. 1a. The Web client (WC) sends the request (1) to the Web server (WS). On the basis of the data identifying the client (2) the decision-making mechanism (DM) determines the intelligent agent (IA) nearest to the client (3). Then the Web server establishes the connection to the agent (4) and gets the information, necessary for the response preparation (5), and generates the response and returns it to the client (6).

The distributed method is shown in Fig. 1b. The idea is based on the assumption that before the client sends the proper query to the server, it contacts with the server earlier, for example by filling out some form needed for sending the query. In that case the client requests the query by means of a two-phase transaction. In the first phase the client (WC) sends the request (1) directly to the server (WS). On the basis of the data identifying the client (2) the decision-making mechanism (DM) in the server determines the intelligent agent (IA) nearest to the client (3). Next, the web server returns to the client a form to be filled out and the nearest intelligent agent's identifier (4). In the second phase the client (WC) sends a filled form but not directly to the server but through its nearest intelligent agent (IA) (5 and 6). The agent plays a role of an intermediary system. The server generates its response which is not yet the final response and sends it to the intelligent agent (IA) (7). The agent shapes the response according to the needs and sends the final response to the client (WC) (8).

The server response shaping made by the agent can be defined in various ways. For example, the agent can filter the results according to the rules specific to local community. If the response from server includes the list of mirrors the agent can determine which mirror has to be used taking into account network and server load.

The response shaping made in the distributed manner via intelligent agents has various benefits. The server is lightened from the load since the agents share the total load needed for the preparation of the final server responses. Then the server may serve much more clients. Agents that are close to their clients can deliver the final responses in much more effective way.

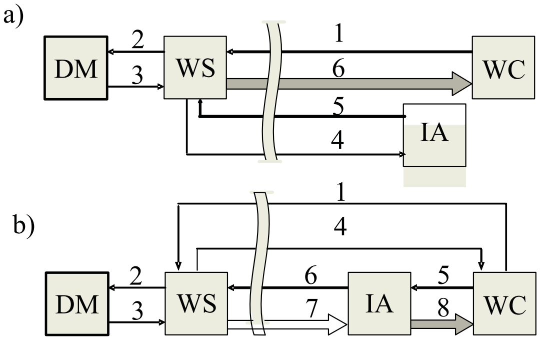


Fig. 1. Centralized (a) and distributed (b) server response shaping

4 Determining the Nearest Agent

Fig. 2 shows the structure of the decision-making mechanism (DM) for the determining the intelligent agent (IA) nearest to the client. DM consists of the distance estimator (DE), knowledge base (KB), comparator (MN) and multiplexer (MR).

The geographic distance is used to determine the nearest agent. We calculate the geographical distance based on the following assumptions. The Earth is an ideal sphere with given radius. The geographical locations of hosts are expressed in the form of spherical co-ordinates of the points on the surface of the sphere. The distance is assigning the length of the shortest arch linking both hosts and lying on the surface of the sphere. The distance between the pair of hosts cannot exceed a half of the perimeter of the sphere. We assume that DE is able to determine the client location (that is the geographical co-ordinates of the host *X* where the client is running) using client's Uniform Resource Identifier l_{WC} (2). The data necessary for that is being taken from the knowledge base (KB) which also collects geographical co-ordinates of all agents (vector L) and their Uniform Resource Identifiers (vector U). All *M* agents are considered and DE estimates for all of them the geographical distance d_m between the client and the *m*-th intelligent agent using formula (1), where m=1,…, *M*. We obtain the vector D of distances between the client location and locations of all agents:

$$\begin{aligned} d_m = \rho \arccos[&\sin(lat_X)\sin(lat_{S_m}) + \\ &\cos(lat_X)\cos(lat_{S_m})\cos(long_{S_m} - long_X)], \end{aligned} \tag{1}$$

where ρ - the earth radius, $long_X, lat_X$ - the longitude and latitude of the client host *X*, and $long_{S_m}, lat_{S_m}$ - the longitude and latitude of *m*-th agent installed on host S_m. The smallest distance is chosen by MN, and *k* is the index of the element in D which has the smallest value. After then the respective *k*-th element of U vector is selected as the identifier u_{k} of the agent chosen by the DM (3).

5 Evaluation

For the purpose of the evaluation we use the Wing system developed at our laboratory [2]. Wing is a network measurement tool that measures end-to-end network path

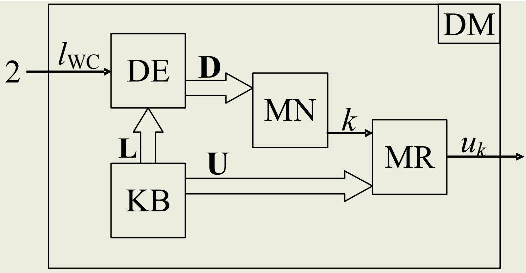


Fig. 2. Decision making mechanism for determining the nearest intelligent agent

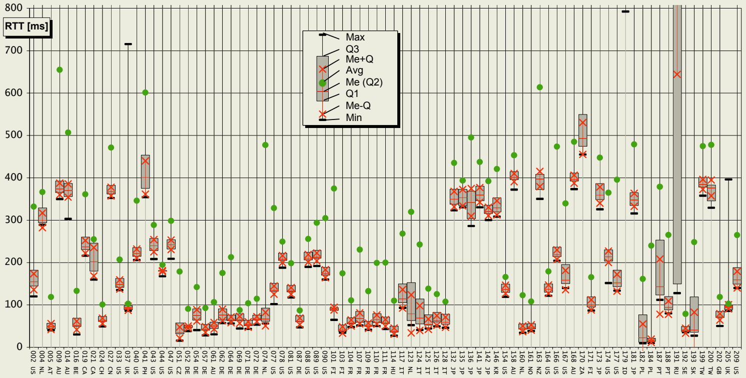


Fig. 3. RTTs for pairs of hosts: WUT host– target host

characteristics at the HTTP layer between the WUT campus network and any target Web site in the Internet. The entire measurement infrastructure is implemented at Wroclaw University of Technology (WUT) side. The Wing works like a sonar location system, accessing periodically a Web resource (a file) from the targeted Web site. The measurements used in this paper were performed between 21 September 2002 and 11 February 2003. The target servers were chosen randomly by the Google search machine. Among a few hundred links found by Google we have chosen 209 direct links to the same resource. After preliminary tests we limited the set of targets to M=83 servers which were active in further measurements. Thus our experiments involved the repeated downloads of resource from 83 different Web servers, ten times a day over 24-hour period. The RTTs were measured between a host placed at WUT campus and 83 target hosts. The geographic localization (longitude, latitude, country, and city) of target server was determined using our host localization service which was developed based on the NetGeo CAIDA's service [12]. The details of experiments can be found in [1, 3]. Fig. 3 presents box & whisker plot of RTT values measured between WUT and target hosts in question. Each box delineates the 25th and 75th percentile of RTT values; the RTT median is represented by the inside line; the ends of whiskers delineate the minimum and maximum values. The average RTT value is represented by a dot. For clarity the values of RTTs are cut off to the values not greater than 800 ms.

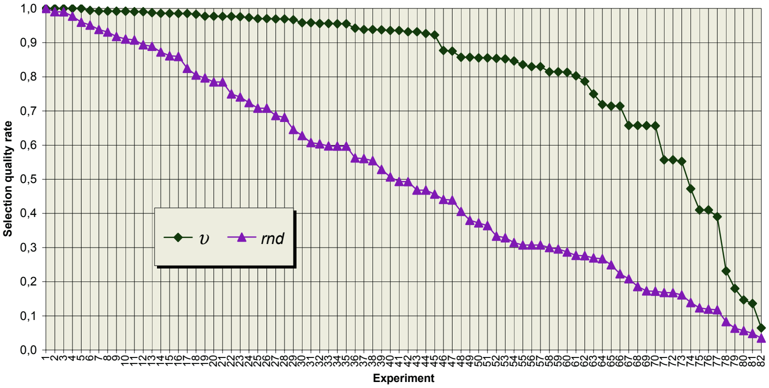


Fig. 4. Selection quality rates

The validation of our method proceeds as follows: We use the RTT median values computed for all 83 pairs of hosts: WUT host – m-th target host, where m=1,2,…, M. We place WS on WUT host for the whole evaluation. We calculate geographic distances from every target host to every other. We also assume that RTT is an ideal distance measure on the Internet. Further evaluation involves M simulation experiments. In m-th experiment among all target hosts one of them is chosen (each time a different host). It is assumed to be a host where the client is placed (host X). Next, the nearest host (i.e. agent) to host X is determined based on the geographic distance between X and other target hosts. This is the nearest agent. After that we compare RTT from WS to chosen nearest agent with RTT from WS to X. If both RTTs report similar results then our selection is good, otherwise is not good. The selection quality rate ξ_m is calculated using (2), where y_m is the RTT median value for the nearest agent and $\bar{y}_m$ is the RTT median value for the client. Fig. 4 shows υ for all experiments – the data is ordered for clarity of presentation, where υ is ξ_m for the nearest agent selection based on geographic distance and *rnd* is the selection quality rate based on a random choice. The average values of selection quality rates are 82% and 51%, respectively.

$$\xi_m = \begin{cases} \dfrac{y_m}{\bar{y}_m}, \text{for } \dfrac{y_m}{\bar{y}_m} \le 1 \\ \dfrac{\bar{y}_m}{y_m}, \text{for } \dfrac{y_m}{\bar{y}_m} > 1 \end{cases} \tag{2}$$

6 Conclusions

In this paper we have studied the problem of choosing the intermediary agent nearest to the web client. The geographic distance based selection has been proposed and evaluated using data sets collected from live Internet. We have shown that the geographic distance for selecting the nearest intermediary agent is accurate and effective. We have achieved an 82% success rate of trials.

References

1. Borzemski L.: Data Mining in Evaluation of Internet Path Performance. In: Innovations in Applied Artificial Intelligence. LNAI, Vol. 3029. Springer-Verlag, Berlin (2004)
2. Borzemski L., Nowak Z.: WING: A Web Probing, Visualization and Performance Analysis Service. In: Web Engineering. LNCS, Vol. 3140. Springer-Verlag, Berlin (2004)
3. Borzemski L., Nowak Z.: An Empirical Study of Web Quality: Measuring the Web from the Wroclaw University of Technology Campus. In: Engineering Advanced Web Applications, Rinton Publishers, Princeton, USA (2004)
4. Dikaiakos M.: Intermediary infrastructures for the World Wide Web. Computer Networks 45 (2004)
5. Francis P., Jamin S., Jin C., Jin Y., Raz D., Shavitt Y., Zhang L., IDMaps: A Global Internet Host Distance Estimation Service, ACM/IEEE Trans. on Networking, vol. 9, no. 5, (2001)
6. Gummadi K., Saroiu S., Gribble S., King. Estimating Latency between Arbitrary Internet End Hosts, Proc. of the SIGCOMM Internet Measurement Workshop (2002)
7. Huffaker B., Fomenkov M., Moore D., claffy kc, Macroscopic analyses of infrastructure: measurement and visualization of Internet connectivity and performance. Proc. of PAM2001 (2001)
8. Huffaker B., Fomenkov M., Plummer D., Moore D., claffy kc, Distance Metrics in the Internet, IEEE International Telecommunications Symposium, ITS (2002)
9. Lakhina A., Byers J., Crovella M., Matta I.: On the Geographic Location of Internet Resources. Proc. of the SIGCOMM Internet Measurement Workshop (2002)
10. Ng E., Zhang H., Predicting Internet Network Distance with Coordinates-Based Approaches, Proc. of the 21st INFOCOMM Conference (2002)
11. Sibson K., Performance Properties of the Web, Matrix.Net (2001)
12. http://www.caida.org

Mining Internet Data Sets for Computational Grids

Leszek Borzemski

Wroclaw University of Technology, Institute of Information Science and Engineering,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
leszek.borzemski@.pwr.wroc.pl

Abstract. Data mining methodology and tools are employed in different application areas. This paper proposes a novel application field for data mining research, namely analysis and long-term forecasting of Internet performance, especially for the needs of Computational Grids. Using data mining the performance problems studied for Internet can be considered from new points of view, and sometimes with better understanding than through applying conventional data analysis methods. This knowledge has been mined by means of professional data mining package in order to build the decision model for advising in further exploitation and usage scheduling of Grid links for a particular time and date. The results show that the data mining can be efficiently used in this research area.

1 Introduction

Nowadays we observe the development of Grid technologies that provide seamless and scalable access to wide-area Internet based distributed resources. Computational Grids are an approach for building dynamically constructed problem solving environments using geographically and organizationally dispersed high performance computing and data handling resources [11]. Grids are used for large-scale problem solving in science and engineering (e.g., high energy physics data analysis, aerospace systems design, cosmology studies, and data mining) [17]. They can be compute-intensive and data-intensive, i.e. Grids developed for speeding up the computations and efficient usage of computing power accessible in local systems in the network, and Grids for very large data handling problem solving. Grids use Internet, preferably high-speed Internet, to communicate between distributed computational resources, transfer data and jobs, and to synchronize computations. The Grids share widely distributed computing resources providing a uniform look and service of distributed systems. There are also industry efforts to combine Grid Services and Web Services (IBM, HP, Microsoft, and Sun) (e.g. see [16]). Commercial users also often demand computing resources and large-scale instruments not available within a single organization or at one location, enabling to develop scalable virtual organization. They need effective services and tools that allow easy sharing of costly and rare computing resources without knowing the specifics of each local environment. They want to match their requirements using the publicly available wide area networking developed around the Internet. Performance network monitoring in Grids is a fundamental problem for them. There are quite a few approaches and systems available for taking measurements and predicting of future performance [e.g. 10, 11, 27]. Most of them

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 268–274, 2005.

are based on the last observations and estimate the future network characteristics in a short-term. One of the most challenging Grid problems is that of scheduling scarce resources such as supercomputers, clusters and large instruments. Such resources have to be co-scheduled for several users and for limited periods of time, so they must be scheduled for a particular time and date. CPU advance reservation scheduling and network bandwidth advance reservation are critical components to the co-scheduling services. If we cannot make network bandwidth reservation we should know the behavior of our network connections to decide which remote resources are to be used in the demanded time period. Therefore we need long-term prediction of network performance.

We can find various data mining applications in science, engineering, business, industry, and medicine [14, 15, 26]. In this work, we characterize a data mining based performance prediction technique that enables the end-user to forecast the log-term network performance behavior and to advise the user in the decision when and which communication line should be used in a particular time and date. Our contribution is to show how data mining methods and tools can be used for analysis of the sets of measured data regarding Internet performance to discover its end-to-end characteristics and to predict its future behavior. Such knowledge about Grid network links may be used for scheduling of future usage of Internet connections for a particular time and date. Data mining aims at finding essential regularities in large data sets. Interest in data mining is motivated by the growth of computerized data sets and by the high potential value of patterns discovered in those data collections.

The rest of the paper is organized as follows. Section 2 presents problem characterization. The example, measurement and mining methodologies are discussed in Section 3. Final remarks appear in Section 4.

2 Problem Characterization

Grids are based on the basic network services – among them there are services for resource discovery, resource co-scheduling, resource brokering and communication services. Many Grids currently use Globus [10] to provide the basic services that characterize and locate resources, initiate, transfer and monitor jobs, provide secure authentication of users, provide uniform access to data, etc. In the Globus Toolkit there is a communications component for providing communication mechanisms for a wide range of communication methods and networks, taking into account network quality of service parameters such as jitter, reliability, latency, and bandwidth.

The Network Weather Service (its functionality is being analogous to weather forecasting) [27] can be used for forecasting the performance of various resource components, including the network itself by sending out and monitoring lightweight probes through the network to the sink destinations at regular intervals. It is intended to be a lightweight, noninvasive monitoring system. This service operates over a distributed set of performance sensors network monitors from which it gathers readings of the instantaneous network conditions. It can also monitor and forecast the performance computational resources. NWS sensors also exist for such components as CPU and disk. NWS runs only in UNIX operating system environments and requires much of installation and administration work. It uses numerical models to generate short-term forecasts of what the conditions will be for a given time frame.

NWS basic prediction techniques are not representative of the transfer bandwidth obtainable for large files (10 MB to 1 GB) and do not support long-term forecasts. New NWS developments address these problems, e.g. [23] shows the technique developed for forecasting long HTTP transfers using a combination of short NWS TCP/IP bandwidth probes and previously observed HTTP transfers, particularly for longer-ranged predictions.

Besides the Grids there is the world of peer-to-peer applications (such as Gnutella [13]) and resilient overlay networks [1]. They are becoming a much portion of Internet traffic. Such P2P application networks are also built among scientific communities (e.g., GriPhyN project [2]). Such initiatives also require well-predictable Internet performance.

Internet performance is extremely difficult to study in an integrated way. It has never been easy to determine whether slow responses are due to either network problems or end-system problems on both sides, i.e. user and server sides, or both. Moreover, because most of these performance problems are transient and very complex in the relationships between different factors that may influence each other, therefore we cannot exactly diagnose and isolate their key sources. Almost 60% latency, as perceived by users at their microscopic level, refers to the end-to-end path between user and Internet host [8]. All these factors may affect ultimate Internet end-to-end performance.

Measurements can merely report the network state at the time of the measurement. When discovered network properties exhibit to be constant over observed life-time span, then the measurements can be used in forecasting the future of system behavior. The assumption of constant network parameters is especially useful for coarser time scale than for fine time scales [28]. The network time delay as measured by the Round-Trip-Time (RTT) is well described as steady on time scale of 10-30 minutes [27]. We also consider RTT as network performance measure. The constancy property of particular network parameter is assumed.

Grid users may need both short-term and long-term network performance forecasts. Short-term forecasting requires instantaneous measuring of network performance. This problem addresses NWS. In long-term forecasting we propose to use historical information. Moreover, sometimes there are some reasons that we are not able measure the network at the particular time. Individual users cannot find long-term performance historical evaluations using the knowledge which is discovered in measurement projects that have been launched on the Internet [e.g., 6, 18, 19, 21]. These projects present their network weather reports that are mainly focused at the whole Internet or a significant part of it, not at particular end-user site, and especially the measurements are mainly performed in the core of Internet.

During exploitation of Internet we can collect measurement data concerning different parameters characterizing Internet performance [5, 6]. For instance, we may gather information about the latency of transmissions for particular end-to-end paths. Such information can be used in construction of prediction performance model. We can passively monitor the Internet paths or organize active experiments. In both situations we usually obtain huge datasets that could not be analyzed using conventional data analysis approach. Moreover, we can expect that the useful information might be hidden in these datasets. But this information usually is not easy to discern using conventional data queries and statistical calculations. In such situation the data mining can be used to discover previously unknown information. Then the resulting information can help make more informed decisions about Internet design and exploitation.

Data mining is a promising area of current research, which can provide important advantages to the users. It is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data which can yield substantial knowledge from data gathered. The ongoing rapid growth of measured data on the Internet has created an immense need for data mining and knowledge discovery methodologies applied to the Internet data.

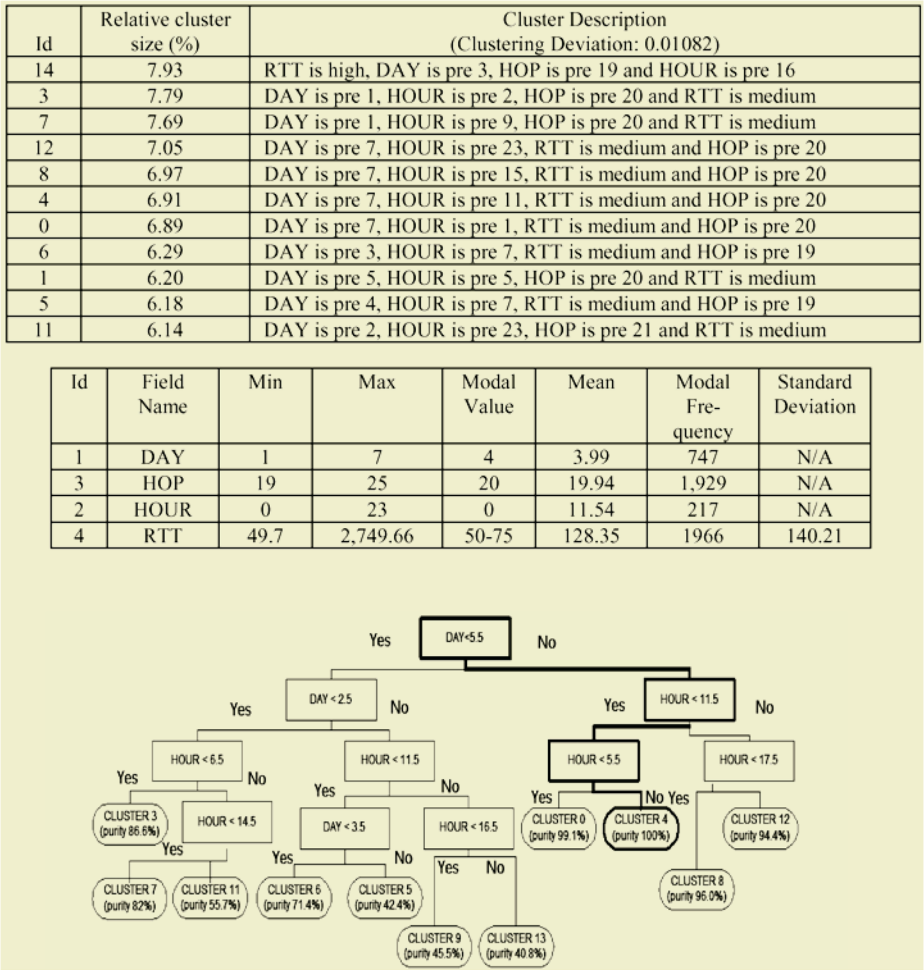
As for today Internet data mining is mainly devoted to Web mining [22] or discovering specific unusual behavior, such as Denial of Service attack [12]. Knowledge discovery in data bases describing Internet performance and functionality begins to be of great interest [9], however the classical statistical data analysis is mostly used in network analysis [1, 2, 3, 5, 6, 10, 18, 20, 27, 28]. We think that data mining methods can be effectively used especially for the analysis of long-term network behavior. Our approach proposes discovering knowledge that may characterize performance behavior of Internet paths, and then making use of this knowledge to advise the user in future network usage.

3 An Example: Creating a Performance Prediction Model

Grid resources have to be co-scheduled for several users and for limited periods of time. In case of the network we may reserve the bandwidth or schedule the usage of the network for a particular time and day. Data mining can help in this by proposing the schedule advice which can be discovered from measurement data. In this section we want to present our approach of employing data mining for long-range predictions. This example is based the research on mining Internet data sets which were collected by us using Trace measurements infrastructure. Data was not measured in a Grid computational infrastructure but the network which was measured is a typical communication subsystem used in Computational Grid.

For data collecting on live Internet we used the *traceroute* network connectivity and routing testing tool [18]. This program tests the route over the network between two destinations on the Internet and lists all the intermediate routers/systems a testing packet must pass through to rich the destination host, as well as the latency between the source host and intermediate systems. Fortunately, it causes low network traffic overhead. Eighteen SLAC servers were selected for measurements [29]. The measurements were performed between Wroclaw and server locations periodically each half an hour and collected in a relational database. Each record includes 21 fields such as: SOURCE and TARGET HOST, TIME STAMP, RTT and HOP. Each host-to-host path is mined individually.

There are two basic data mining functions which we propose to consider here: classification and clustering. Classification assigns data to predefined categories (classes) based on predefined rules. Clustering is similar to classification in that different concept categories called clusters are identified through analysis of characteristics of the data using some proximity measures, but there is no predefined clusters. The clusters are generated through patterns identified in the data.



Id	Relative cluster size (%)	Cluster Description (Clustering Deviation: 0.01082)
14	7.93	RTT is high, DAY is pre 3, HOP is pre 19 and HOUR is pre 16
3	7.79	DAY is pre 1, HOUR is pre 2, HOP is pre 20 and RTT is medium
7	7.69	DAY is pre 1, HOUR is pre 9, HOP is pre 20 and RTT is medium
12	7.05	DAY is pre 7, HOUR is pre 23, RTT is medium and HOP is pre 20
8	6.97	DAY is pre 7, HOUR is pre 15, RTT is medium and HOP is pre 20
4	6.91	DAY is pre 7, HOUR is pre 11, RTT is medium and HOP is pre 20
0	6.89	DAY is pre 7, HOUR is pre 1, RTT is medium and HOP is pre 20
6	6.29	DAY is pre 3, HOUR is pre 7, RTT is medium and HOP is pre 19
1	6.20	DAY is pre 5, HOUR is pre 5, HOP is pre 20 and RTT is medium
5	6.18	DAY is pre 4, HOUR is pre 7, RTT is medium and HOP is pre 19
11	6.14	DAY is pre 2, HOUR is pre 23, HOP is pre 21 and RTT is medium

Id	Field Name	Min	Max	Modal Value	Mean	Modal Frequency	Standard Deviation
1	DAY	1	7	4	3.99	747	N/A
3	HOP	19	25	20	19.94	1,929	N/A
2	HOUR	0	23	0	11.54	217	N/A
4	RTT	49.7	2,749.66	50-75	128.35	1966	140.21

Fig. 1. (a) Characteristics of clusters, pre=predominantly; (b) decision tree

We propose combine these functions in a two-step operation for creation of a prediction model [6]. First, using the clustering function the rules of grouping data with similar properties are discovered. For finding the groups of similar measurements in our input mining database we used the neural clustering algorithm, which employs a Kohonen Feature Map neural network [25]. However, we would like to provide the information about the performance result it would be achieved at a particular time and day in the future. Therefore the decision tree was build based on clusters that were identified at previous step. The resulting tree can be used as the decision-making model in advising the user how to use Internet.

The description of clusters in shown in Fig. 1a. Then using the tree-induction algorithm we build the decision tree. Initially, the tree had 47 nodes and depth 12. Such tree would be ineffective in use, so it was pruned. Fig. 1b shows the decision tree after the pruning of some nodes. The purity in leaf node indicates the percentage of correctly predicted records in that node. The model was tested with known classes and showed 91% of correct classifications.

4 Final Remarks

Data mining consists of the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. Data mining algorithms analyze of the data in large databases to identify trends, similarities, and patterns to support decision making. Data mining models can be predictive and are often used for forecasting of future behavior of the system under consideration.

This paper characterized how we can apply data mining techniques for analysis of Internet performance e.g. for the needs of Computational Grids. We have demonstrated how data mining functions, namely clustering and classifying; can be used to discover the rules that may be used in long-range forecasting of performance of end-to-end Internet path. An example based on real measurement data showed that the resulting decision tree may advice the user in deciding how to use the particular Internet link. Such knowledge about links could be used for scheduling of future usage of Internet connections at a particular date and time.

References

1. Andersen D., Balakrishnan H., Kaashoek F., Morris R.: Resilient Overlay Networks. In: Proc. of 18th ACM Symp. on Operating Systems Principles (2001)
2. Avery, P. and Foster, I.: The GriPhyN Project: Towards Petascale Virtual-Data Grids. GriPhyN TR2001-14, http://www.griphyn.org (2001)
3. Ballintijn G., Van Steen M., Tanenbaumn A. S.: Characterizing Internet Performance to Support Wide-Area Application Development. Operating Systems Review, 34 (4) (2000)
4. Baragoin C., Andersen C.M., Bayerl S., Bent G., Lee J., Schommer C.: Mining Your Own Business in Telecoms Using DB2 Intelligent Miner for Data. SG24-6273-00 (2002)
5. Barford P., Bestavros A., Byers J., Crovella M.: On the Marginal Utility of Network Topology Measurements. In: ACM SIGCOMM Internet Measurement Workshop (2001)
6. Borzemski L.: Data Mining in Evaluation of Internet Path Performance. In: Innovations in Applied Artificial Intelligence. LNAI, Vol. 3029. Springer-Verlag, Berlin (2004)
7. Cabena P., Choi H. H., Kim I. S., Otsuka S., Reinschmidt J., Saarenvirta G.: Intelligent Miner for Data Application Guide. IBM Redbooks, SG24-5252-00 (1999)
8. Cardellini V., Casalicchio E., Colajanni M., Yu P.S.: The State of the Art in Locally Distributed Web-Server Systems. ACM Computing Surveys, Vol. 34, No. 2, June (2002)
9. Faloutsos M., Faloutsos Ch.: Data-Mining the Internet: What We Know, What We Don't, and How We Can Learn More. Full day Tutorial ACM SIGCOMM 2002 Conference, (2002)
10. Foster I., Kesselman C.: Globus: A Metacomputing Infrastructure Toolkit, Intl J. Supercomputer Applications, 11(2), (1997)
11. Foster I., Kesselman C. (Eds.): The Grid: Blueprint for a New Computing Infrastructure, Second Edition, Morgan Kaufmann, Elsevier (2003)
12. Garofalakis M., Rastogi R.: Data Mining Meets Network Management: The NEMESIS Project., Proc. of DMKD'2001 (2001)
13. http://www.gnutellahosts.com.

14. Grossman R. L., Kamath Ch., Kegelmeyer P., Kumar V., Namburu R. R. (Eds.): Data Mining for Scientific and Engineering Applications. Kluwer Academic Publishers, Boston, Dordrecht London (2001)
15. Han J., Kamber M.: Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor, Morgan Kaufmann Publishers, San Francisco (2000)
16. IBM developerWorks: Developing Grid Computing Applications and Web Services, http://www-106.ibm.com/developerworks/
17. Johnston W. E.: Computational and Data Grids in Large-Scale Science and Engineering, LBNL and NASA Ames Research Center, Meeting of the Japanese National Research Grid Initiative project, Tokyo, Japan, http://www-itg.lbl.gov/~wej/ (2003)
18. Luckie M. J., McGregor A. J., and Braun H.-W.: Towards Improving Packet Probing Techniques. In: ACM SIGCOMM Internet Measurement Workshop, San Francisco, CA (2001)
19. Mogul J., Clarifying the Fundamentals of HTTP, In: Proc. of WWW11 Conference (2002)
20. Saroiu S., Gummadi K. P., Gribble S. D.: King: Estimating Latency between Arbitrary Internet End Hosts, In: SIGCOMM Internet Measurement Workshop, (2002)
21. Saroiu S., Gummadi K. P., Dunn R. J., Gribble S. D., Levy H. M.: An Analysis of Internet Content Delivery System. In: Proc. of the Fifth Symposium on Operating Systems Design and Implementation (OSDI 2002) (2002)
22. Srikant R., Yang Y.: Mining Web Logs to Improve Website Organization. In: Proc. of WWW10 Conference (2001)
23. Swany, M., Wolski R.: Multivariate Resource Performance Forecasting in the Network Weather Service. In: Proceedings of the IEEE/ACM SC2002 Conference (2002)
24. Tsuru, M., Oie Y.: Introduction to the Network Tomography. GENESIS Technical Report IEICE Tech. Rep., IN2001-106 (2001)
25. Using Intelligent Miner for Data. V8 Rel. 1, IBM Redbooks, SH12-6394-00 (2002)
26. Wang M., Madhyastha T., Chan N.H., Papadimitriou S., Faloutsos C.: Data Mining Meets Performance Evaluation: Fast Algorithms for Modeling Bursty Traffic. In: Proc. of 18th International Conference on Data Engineering (2002)
27. Wolski, R.: Dynamically Forecasting Network Performance Using the Network Weather Service. Cluster Computing, (1998)
28. Zhang Y., Duffield N., Paxson V., Shenker S.: On the Constancy of Internet Path Properties. In: ACM SIGCOMM Internet Measurement Workshop (2001)
29. http://www.slac.stanford.edu

Towards Integration of Web Services into Agents for Biological Information Resources

In-Cheol Kim and Hoon Jin

Department of Computer Science, Kyonggi University
Suwon-si, Kyonggi-do, 442-760, South Korea
{kic,jinun}@kyonggi.ac.kr

Abstract. Through integration of Web services into agents, we can obtain enhanced functionality and credibility. If agents could access and use Web services, the agents might offer them to other agents as extended services. Additionally, if Web service clients and servers could access and use agent services, agent developers would be able to offer the benefits of agent services to a Web service environment. In this paper, we propose *Agent Service to Web Service Gateway* as a means for increasing availability and usefulness of legacy biological information resources. Specifically, we illustrates as an example Protein Data Base(PDB), which is one of the most important resources for structural proteomics research. We explain the agentification of PDB using a wrapper and then describe how to provide the corresponding Web service through the service gateway in detail.

1 Introduction

Web services are a developing technology stack for representing services and service access in the Internet world. Since Web services tolls are already being developed and deployed by companies, this can provide a number of additional services that can enrich agents with real content. Through integration of Web services into agents, we can obtain the following advantages: If agents could access and use Web services, the agents might offer them to other agents as extended services. Additionally, if Web service clients and servers could access and use agent services, agent developers would be able to offer the benefits of agent services to a Web service environment. In this paper, we propose *Agent Service to Web Service Gateway* as a means for increasing availability and usefulness of legacy biological information resources. Specifically, we illustrates as an example Protein Data Base(PDB), which is one of the most important resources for structural proteomics research. We explain the agentification of PDB using a wrapper and then describe how to provide the corresponding Web service through the service gateway in detail.

2 Biological Information Resources

Biologists are currently generating massive amount of raw data. Outside of the largest, high profile projects such as the Human Genome Project, most of this raw data must be analyzed through the application of various computer programs and searches

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 275–281, 2005.

of various public Web databases. Thus huge databases are being populated with gene and protein data. From a computer science perspective, several problems have arisen.

- Information is available from many distinct locations
- Information content is heterogeneous
- New types of analysis and sources of data are appearing constantly
- Biologists wish to both make their findings widely available, yet retain control over the data

These features make a multi-agent approach particularly attractive.

3 Integrating Web Services into Agents

In an open and distributed agent-based environment, the need of standard mechanisms is crucial to ensure interoperability of distinct autonomous systems. FIPA, which is a standard organization about agent technology, has announced standards for harmonic interacting between heterogeneous agents and supporting services about agent communication, message transporting mechanism, and service directory, etc. According to it, agents communicate with each other in the pre-defined ACL(Agent Communication Language), and the services provided by them should be registered to a service registry called DF(Directory Facilitator). In other words, the DF agent plays a role as yellow page. Therefore in order to request a certain kind of the service from others, an agent can look up the service directory. In a FIPA-compliant agent platform, the AMS(Agent Management System) agent plays a role as white page as well as a manager of each agent's life cycle. On the other hand, Web services are emerging as a new breed of Web application. They are self-contained, self-describing, modular applications that can be published, located, and invoked across the Web. Where the current Web enables users to connect to applications, the Web services architecture enables applications to connect to other applications. Web service technology using XML-based protocols like SOAP, WSDL, and UDDI provides high interoperability. SOAP is W3C's recommended XML-data transport protocol, used for data exchange over web-based communication protocols such as HTTP. WSDL is the W3C recommended language for describing the service interface. UDDI defines a registry service for web services. Web services may be registered with a UDDI registry, which can subsequently be queried by other users and services. Through integration of Web services into agents, we can obtain enhanced functionality and credibility. If agents could access and use Web services, the agents might offer them to other agents as extended services. Additionally, if Web service clients and servers could access and use agent services, agent developers would be able to offer the benefits of agent services to a Web service environment.

Figure 1 represents *Web Service and Agent Service Model* proposed by the Agentcities Web service Working Group. If observed carefully, it can be divided two categories, there are one case which wants to be supported with Web service through agent client and another case which want to be supported with agent service through web client. In the fore one, the service description which agent want to use must be operated after registered on Web service by service provider, and the service contents are described as WSDL(Web Service Description Language) form. WSDL document registered on UDDI is translated to message form for agent service by agent service-to-Web service gateway, and through translated message moved to agent client, finally agent client can ask and receive based on the service content described. In the post one, the service description which web client want to be supported must be made preliminary and registered on the DF agent, if asked by web client, it is supplied to one after translated to WSDL by agent service-to-Web service gateway. Then web client can ask and receive according to the service contents described. In this time, agents who are used in both should be allowed to FIPA agent's needs. While that is, WSAG(Web Services Agent Integration) system which we are willing to use, belongs to the post one and is the gateway system developed for supporting with agent service to Web service. It is composed with Axis, JADE, gateway controller, and gateway agent entirely. Axis is, as a SOAP engine, is a software component managing Web service actually. JADE is situated in WSAG, supporting the operating environment for gateway agent. Gateway controller manages the whole gateway system. Gateway agent has roles between agent service part and Web service one.

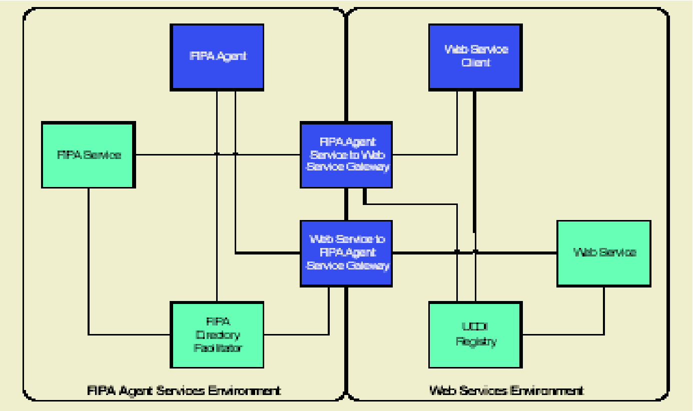

FIPA Agent
Web Service Client
FIPA Service
FIPA Agent Service to Web Service Gateway
Web Service to FIPA Agent Service Gateway
Web Service
FIPA Directory Facilitator
UDDI Registry
FIPA Agent Services Environment
Web Services Environment

Fig. 1. Web Services and Agent Service Reference Model

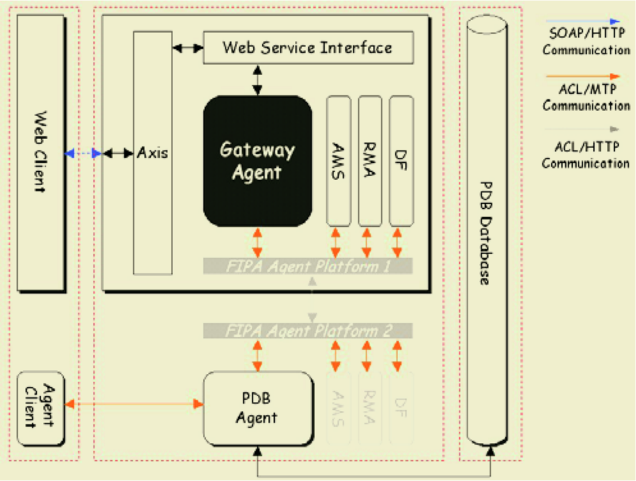


Fig. 2. Architecture of the PSIS

4 Protein Structural Information System

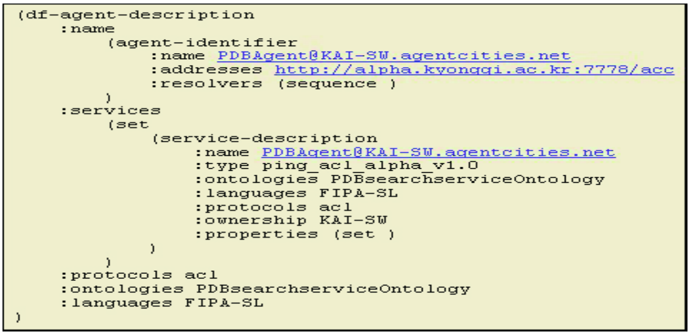
Our PSIS(Protein Structural Information System) is a multi-agent system for providing protein structural information. It consists of the components described in figure 2. The three major components of the PSIS are the PDB database part (right), the service creation part (mid), and the service client part (left). The service creation part is again composed of the PDB agent and the gateway agent. The PDB agent is a kind of wrapper program that makes the PDB database a FIPA-compliant agent. The gateway agent translates the PDB agent service into a Web service. Through the PDB agent and the gateway agent, the PSIS system provides two distinct types of services simultaneously: agent services and Web services. The agent services and the Web services provided by the PSIS are registered with a DF agent and a UDDI registry, respectively. And then the client agent and the Web service client program can request a service independently from the PSIS after discovering it from the DF and the UDDI registry.

The services provided by the PSIS are ID-based search, keyword-based search, and sequence-based search on the PDB database. Table 1 lists up the services along with their input parameters and the resulting outputs. The table classifies the services based on their query types(ID/keyword/sequence), search methods(exact/full), result types(file/sequence/abstract information/list), and file formats(pdb/mmcif/xml). Three different file formats for representing protein structures are available in our PSIS: the original PDB format, the MMCIF format, and the XML-based format. As mentioned above, these search services can be provided simultaneously both in the form of agent service and Web service.

Table 1. PDB Search Services

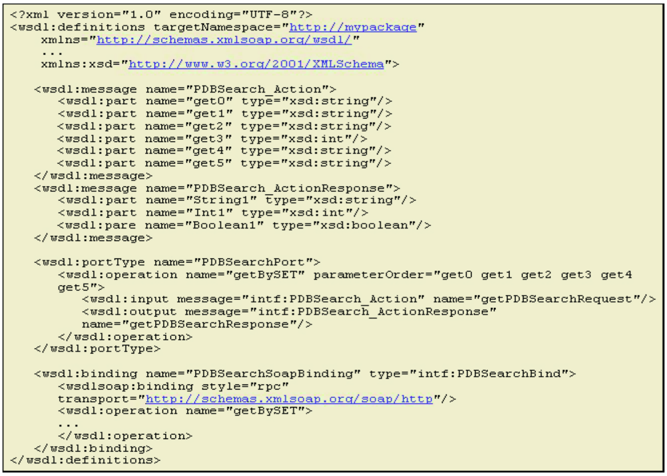
- ● ID based search
  - ■ Exact search(target: single)
    - (result: file)
      - − (format: pdb/mmcif/xml)
    - (result: sequence)
    - (result: abstract information)
  - ■ Full search(target: multiple)
    - (result : list)
- ● Keyword based search
  - ■ Exact search (target: single)
    - (result: file)
      - − (format: pdb/mmcif/xml)
    - (result : sequence)
    - (result : abstract information)
  - ■ Full search (target: multiple)
    - (result : list)
- ● Sequence based search
  - ■ Full search(target: multiple)
    - (result : list)

Fig 3 and Fig 4 shows the agent service description and the Web service description of our PSIS, respectively. While Fig 3 represents the DF description for the agent service, Fig 4 represents the WSDL document for the Web service. Based upon these service descriptions, the service client programs can be built and executed to send a proper request message to the PSIS system. Fig. 5 shows an example of SOAP service request messages from a remote Web service client. The client requests an ID-based exact search on the PDB database. This SOAP request message will be translated into an ACL message by the gateway agent, and then delivered to the PDB agent. Through interactions with the remote PDB database, the PDB agent formulates the search result into an ACL response message, and then transfers it to the gateway agent. Then the gateway agent translates the ACL message into a SOAP response message in reverse and returns it to the Web service client.



```
(df-agent-description
    :name
        (agent-identifier
            :name PDBAgent@KAI-SW.agentcities.net
            :addresses http://alpha.kyonggi.ac.kr:7778/acc
            :resolvers (sequence )
        )
    :services
        (set
            (service-description
                :name PDBAgent@KAI-SW.agentcities.net
                :type ping_acl_alpha_v1.0
                :ontologies PDBsearchserviceOntology
                :languages FIPA-SL
                :protocols acl
                :ownership KAI-SW
                :properties (set )
            )
        )
    :protocols acl
    :ontologies PDBsearchserviceOntology
    :languages FIPA-SL
)
```

Fig. 3. The Agent Service Description



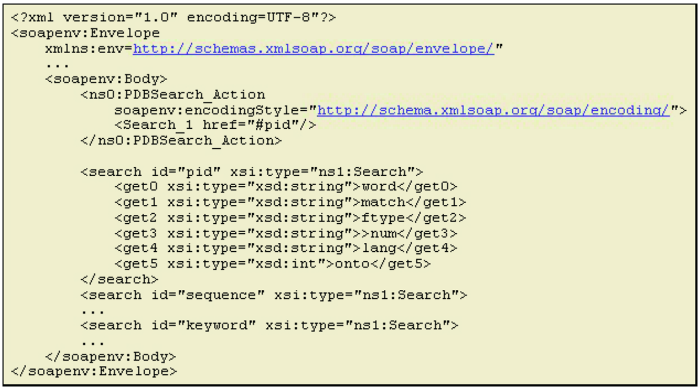
```
<?xml version="1.0" encoding="UTF-8"?>
<wsdl:definitions targetNamespace="http://mypackage"
    xmlns="http://schemas.xmlsoap.org/wsdl/"
    ...
    xmlns:xsd="http://www.w3.org/2001/XMLSchema">

   <wsdl:message name="PDBSearch_Action">
      <wsdl:part name="get0" type="xsd:string"/>
      <wsdl:part name="get1" type="xsd:string"/>
      <wsdl:part name="get2" type="xsd:string"/>
      <wsdl:part name="get3" type="xsd:int"/>
      <wsdl:part name="get4" type="xsd:string"/>
      <wsdl:part name="get5" type="xsd:string"/>
   </wsdl:message>
   <wsdl:message name="PDBSearch_ActionResponse">
      <wsdl:part name="String1" type="xsd:string"/>
      <wsdl:part name="Int1" type="xsd:int"/>
      <wsdl:pare name="Boolean1" type="xsd:boolean"/>
   </wsdl:message>

   <wsdl:portType name="PDBSearchPort">
      <wsdl:operation name="getBySET" parameterOrder="get0 get1 get2 get3 get4
      get5">
         <wsdl:input message="intf:PDBSearch_Action" name="getPDBSearchRequest"/>
         <wsdl:output message="intf:PDBSearch_ActionResponse"
         name="getPDBSearchResponse"/>
      </wsdl:operation>
   </wsdl:portType>

   <wsdl:binding name="PDBSearchSoapBinding" type="intf:PDBSearchBind">
      <wsdlsoap:binding style="rpc"
      transport="http://schemas.xmlsoap.org/soap/http"/>
      <wsdl:operation name="getBySET">
      ...
      </wsdl:operation>
   </wsdl:binding>
</wsdl:definitions>
```

Fig. 4. The Web Service Description in WSDL

The developing environment for PSIS is like the followings:

- H/W: Intel Pentium4 dual 2.0, 1G memory
- S/W: SDK1.42, AXIS1.1, Tomcat Server4.1, jUDDI4.0, JADE3.0b, WSAG1.0

Figure 6 shows an example search with the PSIS system. It represents the result of an ID-based exact search service with the protein ID of "*1aeg*". The requested file format was the original PDB format. The received protein structural information is being displayed graphically using a local viewer embedded in the PSIS system.

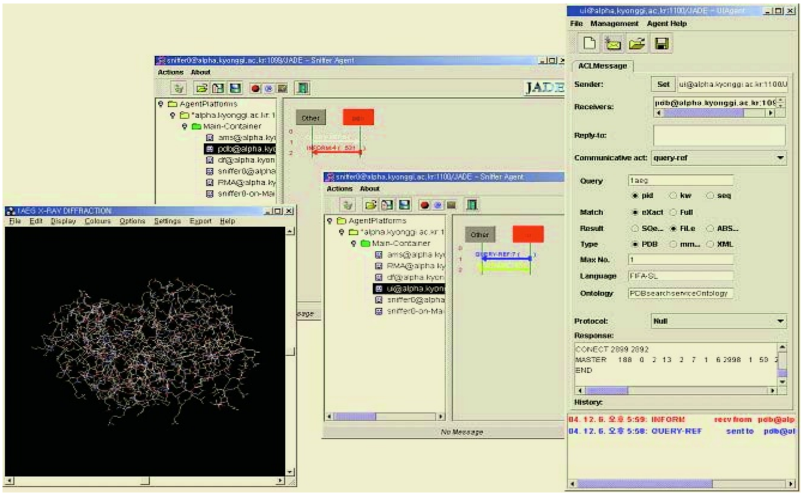


```
<?xml version="1.0" encoding=UTF-8"?>
<soapenv:Envelope
    xmlns:env=http://schemas.xmlsoap.org/soap/envelope/"
    ...
    <soapenv:Body>
        <ns0:PDBSearch_Action
            soapenv:encodingStyle="http://schema.xmlsoap.org/soap/encoding/">
            <Search_1 href="#pid"/>
        </ns0:PDBSearch_Action>

        <search id="pid" xsi:type="ns1:Search">
            <get0 xsi:type="xsd:string">word</get0>
            <get1 xsi:type="xsd:string">match</get1>
            <get2 xsi:type="xsd:string">ftype</get2>
            <get3 xsi:type="xsd:string">>num</get3>
            <get4 xsi:type="xsd:string">lang</get4>
            <get5 xsi:type="xsd:int">onto</get5>
        </search>
        <search id="sequence" xsi:type="ns1:Search">
        ...
        <search id="keyword" xsi:type="ns1:Search">
        ...
    </soapenv:Body>
</soapenv:Envelope>
```

Fig. 5. A SOAP Request Message

5 Conclusions

In this paper, we proposed "*Agent Service to Web Service Gateway*" as a means for increasing availability and usefulness of legacy biological information resources.



Sender: Set ui@alpha.kyonggi.ac.kr:1100U

Receivers: pdb@alpha.kyonggi.ac.kr:109

Reply-to:

Communicative act: query-ref

Query 1aeg

● pid ○ kw ○ seq

Match ● eXact ○ Full

Result ○ SOe... ● FiLe ○ ABS...

Type ● PDB ○ mm... ○ XML

Max No. 1

Language FIFA-SL

Ontology PDBsearchserviceOntology

Protocol: Null

Response:

```
CONECT 2899 2892
MASTER    188   0   2  13   2   7   1   6 2998   1  59  2
END
```

History:

```
04. 12. 6. 오후 5:59: INFORM       recv from  pdb@alp
04. 12. 6. 오후 5:58: QUERY-REF      sent to   pdb@al
```

Fig. 6. A Screenshot of the PSIS

Specifically, we illustrated as an example Protein Data Base(PDB), which is one of the most important resources for structural proteomics research. We explained the agentification of PDB using a wrapper and then describe how to provide the corresponding Web service through the service gateway in detail.

Acknowledgement

This research was supported by Kyonggi University Overseas Research Grant 2003.

References

1. Hoon Jin, In-Cheol Kim: Plan-based Coordination of Multi-Agent System for Protein Structure Prediction, Proceedings of the 26th Internal Conference on AI, Simulation and Planning, LNCS/LNAI 3378, AIS2004, 4-6 October 2004, South of Korea (2004).
2. Dominic Greenwood, Monique Calisti: Engineering Web service – Agent Integration, Systems, Cybernetics and Man Conference, 10-13, October, The Hague, Netherlands. (2004).
3. Hoang Pham Huy, Takahiro Kawamura, Tetsuo Hasegawa: From Web Browsing to Web service - Fertilizing Agent Environment, AAMAS'2003, Workshop on Web services and Agent-based Engineering, july. (2003).
4. Michael Y. Galperin: The Molecular Biology Database Collection: 2004 update, Nucleic Acids Research, Vol.32, Database issue. (2004).
5. PDB Database, http://www.rcsb.org/pdb/.
6. Agentcities Network, http://www.agenticities.org/.

Learning Within the BDI Framework: An Empirical Analysis

Toan Phung, Michael Winikoff, and Lin Padgham

RMIT University, School of Computer Science and IT, Melbourne, Australia
{tphung,winikoff,linpa}@cs.rmit.edu.au

Abstract. One of the limitations of the BDI (Belief-Desire-Intention) model is the lack of any explicit mechanisms within the architecture to be able to learn. In particular, BDI agents do not possess the ability to adapt based on past experience. This is important in dynamic environments since they can change, causing methods for achieving goals that worked well previously to become inefficient or ineffective. We present a model in which learning can be utilised by a BDI agent and verify this model experimentally using two learning algorithms.

1 Introduction

Intelligent Agents are a new way to develop software applications. They are an amalgam of Artificial Intelligence (AI) and Software Engineering concepts that are highly suited to domains that are inherently complex and dynamic [1, 2]. Agents are autonomous in that they are able to make their own decisions. They are situated in an environment and are reactive to this environment yet are also capable of proactive behaviour where they actively pursue goals. BDI (Belief Desire Intention) agents are one popular type of agents which support complex behaviour in dynamic environments [3, 4].

BDI agents use plans to achieve goals, based on the current environment. When a BDI agent encounters a problem where it can not complete the current plan, it will stop executing that plan, re-assess the situation based on the updated environment and select a new plan from a plan library. This provides a level of adaptivity to the changing world. However it does not provide any adaptation based on past experience. Such an ability can be important in dynamic environments which may change in ways not foreseen by the developer, causing methods for achieving goals that worked well previously to become inefficient or ineffective. Our work aims to improve BDI agents by introducing a framework that allows BDI agents to alter their behaviour based on past experience, i.e. to learn.

We have chosen a fire fighting scenario as our experimental domain. This system simulates a city that has been affected by fire and will allow us to learn answers to questions such as *“Based on past experience, what’s the best fire extinguisher to use now?”* We do not develop new learning techniques, rather, the contribution of this paper is to propose a model for integrating learning into BDI agents, and to experimentally validate that this model allows agents to improve their performance over time.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 282–288, 2005.

2 Background

2.1 The BDI Agent Architecture and JACK

The Belief-Desire-Intention (BDI) [3] model is based on philosophical work by Bratman, which stresses the importance of *intentions*, defining the (human) agent's current approach, as critical in intelligent behaviour, as well as *beliefs* and *desires*. The computational model of agency developed by Rao and Georgeff [4] based on Bratman's work focusses on (software) agents which are situated in an environment, receiving stimulus in the form of *events* and acting based on *plans* in a plan library.

JACK[1] is a Java-based intelligent agent toolkit used to implement BDI agents. There are four main components to a JACK system: agents, events, plans and beliefsets. When a JACK agent receives an event, which may correspond to a goal, it will refer to its *plan library*. Plan libraries act as a repository of plans. Plans consist of (i) a *trigger* which indicates which event they are relevant to; (ii) a *context condition* which describes the situation in which they are applicable; and (iii) a *plan body*. The plan body may contain both *sub goals* and actions. There may be multiple plans associated with any given goal or event. If a plan fails during execution, the agent checks to see whether other plans are applicable. Beliefsets can be viewed as relational databases, i.e. sets of tuples.

2.2 Inductive Logic Programming and Alkemy

Inductive logic programming (ILP) is a means of computationally achieving induction [5]. Induction can be defined as: given a set of positive examples, a set of negative examples, some background knowledge and a hypothesis language, find a predicate definition represented in the hypothesis language such that all positive examples and none of the negative examples are described.

Alkemy [6] is a symbolic inductive learner written in C++. It uses Inductive Logic Programming to produce a decision tree (see figure 1). Each node in the decision tree generated by Alkemy contains a higher order function that takes an *Individual* and returns a Boolean. For example, consider the root node in figure 1. The composition `projIntensity . eqHigh` is the function that takes an Individual and returns true iff its Intensity is High. The whole root node expression denotes a function that takes an Individual and returns true iff its Intensity is High and its Weather is Hot.

3 Learning in the BDI Framework

Our framework consists of four major components: the JACK system, the *Learning Formatter*, the *Learning Component* and the *Knowledge Extractor*. Figure 2 shows our conceptual model.

The flow of information begins with the *JACK agent*. This agent stores its experiences in the *History* beliefset. When enough history accumulates, *Learning Formatter* converts the History and *Background Knowledge* (provided by the agent designer) into an input suitable for the *Learning Component*. Learnt data is returned and converted into

[1] http://www.agent-oriented.com

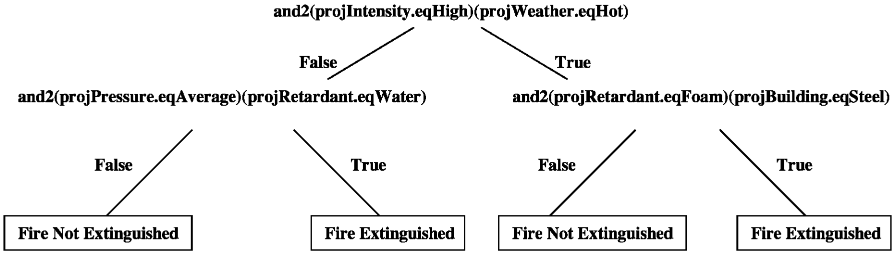

and2(projIntensity.eqHigh)(projWeather.eqHot)
False
True
and2(projPressure.eqAverage)(projRetardant.eqWater)
and2(projRetardant.eqFoam)(projBuilding.eqSteel)
False
True
False
True
Fire Not Extinguished
Fire Extinguished
Fire Not Extinguished
Fire Extinguished

Fig. 1. Higher Order Function Tree

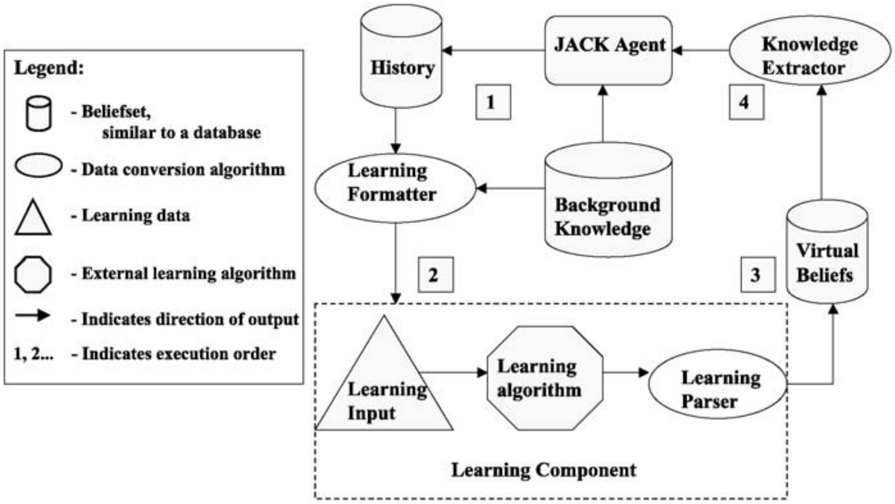


Fig. 2. BDI Learning Model

Virtual Beliefs by the *Learning Parser* which translates Alkemy's textual output into a binary tree. These virtual beliefs are queried by the *Knowledge Extractor*, allowing the agent to reason historically.

The History stored by the agent is a set of tuples containing the *state of the fire*, the *outcome* and a *retardant*. For example, the History tuple ⟨windy, concrete, high, success, water⟩, represents that it was a windy day when water was successfully used to extinguish a concrete building burning with a high intensity. Actually, in order to experiment with different search space sizes the *state of the fire* varies from 3 to 11 elements.

The operation of the *Knowledge Extractor* involves the following steps: (1) Estimate the accuracy of the tree produced by Alkemy; (2) If the accuracy is "good enough" (see below) then use the recommendation produced by the tree, else explore.

Estimating the accuracy of the decision tree is done by checking the tree's predictions against the actual outcome for all of the recent fires that the agent has fought which have not yet been used for learning. This gives a number between 0 and 1. For example, if there are 37 recent fires that have not yet been learned from, and for 32 of them the decision tree correctly predicts the outcome, then the estimated accuracy of the Alkemy tree is $32 \div 37 \approx 0.865$.

Producing a recommendation from the Alkemy decision tree is done as follows. First, the Knowledge Extractor scans the higher order function tree to see what values exist for the *retardant* variable. If none are found, then the agent has had no relevant prior experience and will return *unknown* or a default value. If values are found, then the Knowledge Extractor will record every unique value[2]. This forms a set *potential* retardants to use. For each potential retardant the Knowledge Extractor uses the decision tree to predict the outcome of using that retardant on the current fire. Those retardants for which the decision tree predicts a successful outcome are retained as the tree's recommendation.

Determining whether the tree's accuracy is "good enough" is done in a number of ways: using a static threshold (e.g. 0.5), using a dynamic threshold with analogous reasoning, or using a dynamic threshold without analogous reasoning. When using a dynamic threshold, the threshold is adjusted up or down by considering the subset of the fires previously encountered which are either the same (the "without analogous reasoning" case) or "similar" (the "with analogous reasoning" case). Adjusting the threshold is done as follows: for each fire that is considered we adjust the threshold up if the fire was successfully fought, and down if it was not successfully fought. The formula used to calculate dynamic thresholding is:

$$\text{threshold} = \text{static threshold} - \frac{\text{successful cases} - \text{failed cases}}{2 \times \text{total cases}}$$

If analogous reasoning (also termed "*Simile*") is used then a previously encountered fire is considered to be "similar" if it was successfully fought and is harder than the current fire (because fighting the current, easier, fire can be assumed to succeed) or if it was unsuccessfully fought and is easier than the current fire. For example, suppose the agent is fighting a fire in *Hot* weather where the building is made of *Wood* and the fire is burning with a *High* intensity. A previously fought fire that was on a *Mild* day, involved a *Steel* building, and was a *Medium* intensity fire is easier than the current fire. If a particular retardant was unsuccessfully used on the previous, easier, fire than the simile algorithm will reason that the retardant in question is probably a bad choice for the current, harder, fire.

Exploration is implemented by subtracting all the previously seen retardants from the full list of known retardants, and selecting a random retardant from the result. If the result is empty then a random previously seen retardant that is not recommended by the tree is chosen.

In addition to using Alkemy, we also experiment with a simpler learning mechanism that simply computes for each retardant its effectiveness:

$$\text{effectiveness} = \frac{\text{successes} - \text{failures}}{\text{total}}$$

The retardant with the highest effectiveness is then selected. There are several variants of this depending on whether one considers all past fires, or only past fires similar to

[2] As well as an additional "none-of-the-above" value

the current fire. Note that this simpler mechanism bypasses the learning component depicted in figure 2, since it only requires the agent's history.

4 Experiments

Experiments were conducted within the fire fighting domain to answer the following questions: (1) how effective are various learning mechanisms on BDI agents? (2) what effects do dynamic thresholding and Simile have on learning? and (3) how is the performance of the agent affected by the size of the search space?

Each experiment involved 40 runs, where a run involved a learning agent fighting 1000 fires using one of five retardant types. The fire fighting agent is given no initial past experiences. Fire states were randomly generated using a random number generator that was initialised with a different seed for each run. Alkemy is invoked every 50 fires. The performance of the agent is measured by the percentage of fires extinguished over a given set of fires.

To determine whether a fire is successfully extinguished, we convert every symbolic fire state into a numeric representation and compare that value to a set of rules. In order to do this, every variable is given its own 'difficulty' score. This represents how 'hard' a particular tuple variable is, thus the difficulty score for a entire fire is the sum of the difficulty scores in the fire tuple. Symbolic-to-numeric conversion is done to allow us to easily vary the complexity of the domain. The complexity of the search space is varied from an initial 1440 possible fire states to 2,304,000 by increasing the number of variables in the fire state from 3 through to 11.

All graphs were plotted with the mean of the 40 experimental runs. Each point on the '% of Fires Extinguished' graph represents the success rate over the most recent 50 fires.

4.1 Discussion

Clearly, learning is beneficial to the agent's performance. For the smaller search space (576,000) the statistical method without Simile does best, followed by the Statistical method with Simile, then Alkemy. For the larger search space (2,304,000) Alkemy outperforms all other learning methods by 6%-10% followed by the Statistical method (with Simile not making much of a difference). Overall, learning in the smaller search space yields a 10% improvement over no learning with a 27% improvement in the larger search space.

Although not represented on the graphs, using dynamic thresholding and Simile did not make a difference to Alkemy's success rate. This is because the decision tree rapidly becomes quite accurate, resulting in the slight adjustments to the threshold made by dynamic thresholding not being significant.

Comparing the different search space sizes, as expected, the statistical method's performance degrades as the search space size increases. However, Alkemy's performance doesn't appear to be significantly affected by the search space, and in fact Alkemy does slightly better in terms of % of fires extinguished when the search space is larger. We intend to investigate this counter-intuitive result further.

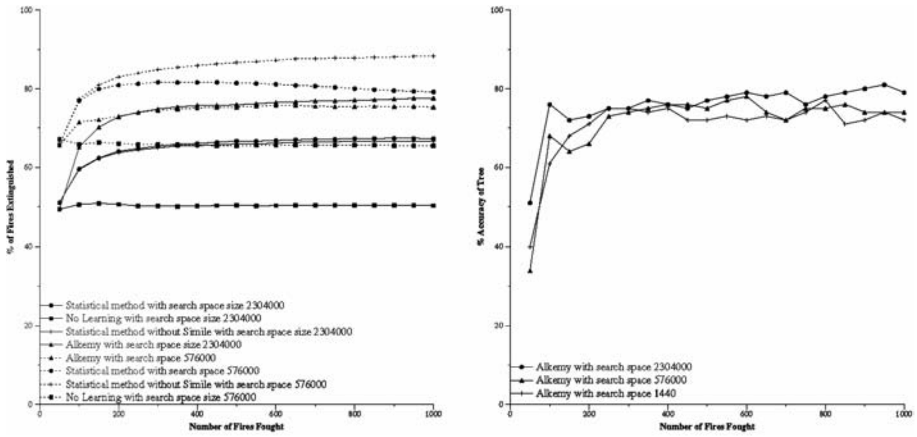


Fig. 3. Experimental Results

With regard to the second graph (right side of figure 3), the accuracy of the Alkemy tree as measured by the agent is quite erratic and never rises above 81%. This may be because we convert symbolic states into numeric values and the fact that Alkemy is a symbolic learner. We intend to explore this further.

Although Alkemy extinguishes more fires than Statistical learning by an average of 10% in the more complex domain, this comes at a time cost 4 times greater than that of Statistical learning. The Statistical method out-performs Alkemy in the simpler domain, highlighting the fact that complex and powerful learners such as Alkemy are not always necessary.

For both search spaces (2,304,000 and 576,000) Alkemy took on average a little over a minute (61-65 seconds) to induce a decision tree from 500 fires.

5 Related Work

Similar systems to what we propose include SOAR [7], a rule based agent system that uses *chunking* to create plans. Chunking is executed whenever *impasses* occur. An impass is when an agent cannot solve a problem. Our model is different in that we learn new information regardless of problems occurring, which allows for exploratory learning.

The Case-Based BDI system in [8] is similar to our model where it considers past cases. They use a *concept hierarchy* to find information on the WWW if no similar cases are found while we assume no additional information sources and hence use *Simile* to reason further on existing information. The notion of ‘easier‘ and ‘harder‘ for case similarity is absent in [8] however their model applies case reasoning on agent beliefs while we do not.

The system proposed by [9] uses a combination of explanation based learning (EBL) and ILP. EBL uses only one past case to generalise a rule while our statistical method considers all past cases. Another difference is our model uses Simile to filter cases before an ILP system is called.

Prodigy [10] is a planning and learning system that implements many learning algorithms including case-based reasoning and induction. However, their work is not based on the BDI framework.

6 Conclusions and Future Work

We have presented a model that introduces learning into the BDI framework. This model allows beliefs to be generalised through inductive learning and statistical tallying. We have developed and experimentally tested, two analogous reasoning algorithms which use contextual and relative reasoning to alter agent behaviour according to past experience.

Currently, the issue of *when* to learn is addressed by means of a numeric threshold on the number of fires fought. This static technique may greatly over/under utilise a potentially expensive[3] learning process and may be improved by considering the frequency of past successes/failures. We also plan to develop a more effective Simile matching scheme.

Acknowledgements

We would like to thank the Smart Internet CRC for their support with this project.

References

1. N. R. Jennings: An Agent-based Approach for Building Complex Software Systems. Communications of the ACM **44(4)** (2001) 35–41
2. Wooldridge, M.: An Introduction To MultiAgent Systems. first edn. John Wiley and Sons Ltd (2002)
3. Bratman, M.E.: Intentions, Plans, and Practical Reason. first edn. Harvard University Press, Cambridge, MA (1987)
4. A. S. Rao, M. P. Georgeff: BDI-Agents: From Theory to Practice. In: Proceedings of the First International Conference on Multiagent Systems. (1995)
5. Muggleton, S.: Inductive Logic Programming. first edn. Academic Press (1992)
6. Kee Siong Ng: Alkemy: A Learning System Based on an Expressive Knowledge Representation. Available from http://users.rsise.anu.edu.au/~kee/Alkemy/ (2004)
7. J.E. Laird, A. Newell, P.S. Rosenbloom: SOAR: An Architecture for General Intelligence. Artificial Intelligence **3** (1987) 1–64
8. C. Olivia, C.F Chang, C.F Enguix, A.K. Ghose: Case-Based BDI Agents: An Effective Approach for Intelligent Search on the World Wide Web. In: AAAI Spring Symposium. (1999)
9. E. Alonso, D. Kudenko: Logic-Based Multi-Agent Systems for Conflict Simulations. In: Proceedings of the 5th UK Workshop on Multi-Agent Systems UKMAS'00. (2000)
10. M. Veloso, J. Carbonell, A. Perez, D. Borrajo, E. Fink, J. Blythe: Integrating Planning and Learning: The PRODIGY Architecture. Journal of Experimental and Theoretical Artificial Intelligence **7(1)** (1995)

[3] In terms of computational resources

How to Make Robot a Robust and Interactive Communicator

Yoshiyasu Ogasawara[1], Masashi Okamoto[1], Yukiko I. Nakano[2],
Yong Xu[3], and Toyoaki Nishida[3]

[1] Graduate School of Information Science and Technology, the University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{yoshiyas,okamoto}@kc.t.u-tokyo.ac.jp
[2] Research Institute of Sci. and Tec. for Society, Japan Sci. and Tec. Agency
Atago Green Hills MORI Tower 18F,2-5-1 Atago, Minato-ku, Tokyo,105-6218, Japan
nakano@ kc.t.u-tokyo.ac.jp
[3] Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
xuyong@ii.ist.i.kyoto-u.ac.jp, nishida@i.kyoto-u.ac.jp

Abstract. One of the key points to make robot a robust and interactive communicator is enabling robot to recognize user's attention behavior and transit among communication situations, so that human and robot can be involved into shared activity. Four communication situations are defined to describe typical situations of interactive human-robot communication. The robot can not only open and close communication, but also adapt to user's behavior passively or actively. We proposed a two-layered architecture of robot system. Implementation of listener robot partly proved effectiveness of the proposal approach.

1 Introduction

New applications for robots bring them into the human environment where they are to serve as helpful assistants to untrained users in the home or office. These applications require robots to be able to perform tasks from natural human instruction and to collaborate with human partners. It is very possible that the robot will become a part of our daily lives in near future. The interaction capability is required for the robots in order to communicate smoothly with human users. Our research concerns how to make a robot interact with people in a human-like natural way. Since the robots' functions can be very complicated, it is difficult for user, especially for novice users, to learn how to use robot and how to convey user's intention to the robot. Since the NLP (Natural Language Processing) technology, which focuses on processing verbal information, is not mature enough to build practical application at present, it is natural and easy for untrained user to learn the usage of robot by nonverbal way instead of traditional way, such as reading piles of manuals.

The object of our research is to build an interactive and adaptive robot system. Firstly, the system should be natural for an untrained user to communicate with and adapt to. Secondly, it should follow social habits to open and close communication with user. Thirdly, the system should focus on recognizing user's attention behavior and generating proper response. Also the system should be able to adjust its own

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 289–295, 2005.

action by adapting to user's indication or interacting with the user. The work of Sidner [1] focuses on evaluating human-robot interaction with a penguin robot. The robot can produce engagement behaviors in collaborating with a human conversational partner. However we prefer to use humanoid robot, because we think it may act better than a pet robot when generating human-like behavior, such as nodding, waving, pointing gesture.

The main contribution of this paper consists of a two-layered architecture of the robot system to satisfy the conditions mentioned above according to a particular communication situation.

2 Communication Situation

It is necessary for a robot to follow social habits, and recognize user's attention behavior, especially nonverbal behavior, so that it can naturally response to user during human-robot communication. Additionally the robot should be able to adjust its own actions by adapting to user's movement. To satisfy those conditions according to a particular communication situation, we propose a two-layered system as a solution. A communication situation is based on some specific conditions, such as time, place and participant, which are involved in an activity. In this paper, communication situation means a situation where some specific conditions occur in shared activity.

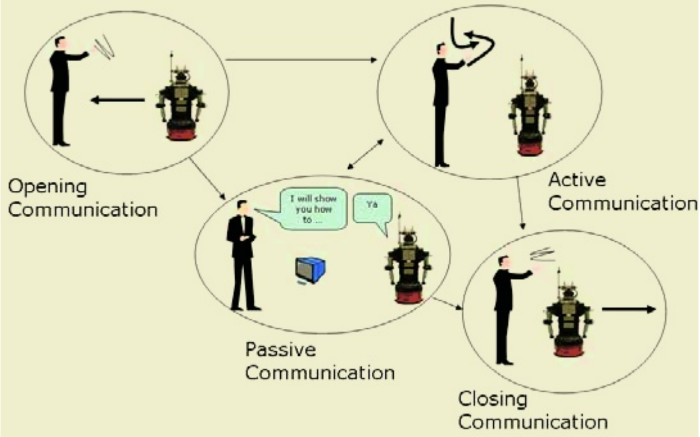


Fig. 1. Communication Situations

We defined two pairs (four types) of communication situations as shown in Fig. 1: opening communication situation, closing communication situation, passive communication situation and active communication situation. We will explain the four types of communication situation respectively.

2.1 Opening Communication Situation

Opening communication situation is literally the first phase of communication. For example, when a human user raises his arm or repeats waving his arms or hands intentionally for several times, the communication is opened. In this case, the robot will observe user's behaviors to estimate whether and when to enter the communication. After approaching to user, the robot can initiate communication by involving user and robot into shared activity. Suppose in a party hall, there is a waiter robot which provide service for customers, such as offer drink, remove empty cups etc. In this case, the robot should be able to recognize human's behavior following social habits. For example, if a customer raises his hand or waving his arm, it may means that he has some requests to call the waiter, and the robot should be able to recognize the movement and give some appropriate response, such as approaching to the customer and ask for his requests, then the communication is opened.

2.2 Closing Communication Situation

When the robot finished a step of task or completed the whole task, the communication procedure should be closed. The robot can estimate whether and when to leave the communication by observing human user's behaviors. It will leave human user and close communication when a specific task has been finished. In the example of waiter robot, if the robot finished removing empty cups, or user turns to other people and does not look at the robot for long time, it may means the task is finished and the communication can be closed.

2.3 Passive Communication Situation

The passive communication situation is the situation when the user explains something to the robot and the robot adapt to user passively and mainly behaves as a listener. In this case, the robot behaves as a passive participant, just like a listener or a student. It observes human's specific behavior, such as explanation and demonstration and gives some proper responses, such as nodding or saying “Ya”, to make the communication go on smoothly.

2.4 Active Communication Situation

In active communication situation, the robot performs actions and adjusts movements by sensing and adapting to user's indication actively. The robot behaves as an active participant who acts according to a human's instructions and adjusts its own movement during the progress of actions by adapting to user's indication or interacting with human user.

2.5 Transition of Communication Situations

The transition of communication situations is also shown in Fig.1. Firstly opening communication is concerned. Secondly the transition between active or passive communication situations will be considered. Lastly closing communication situation occurs.

3 The Architecture of the System

The architecture of the system is shown in Fig.2.

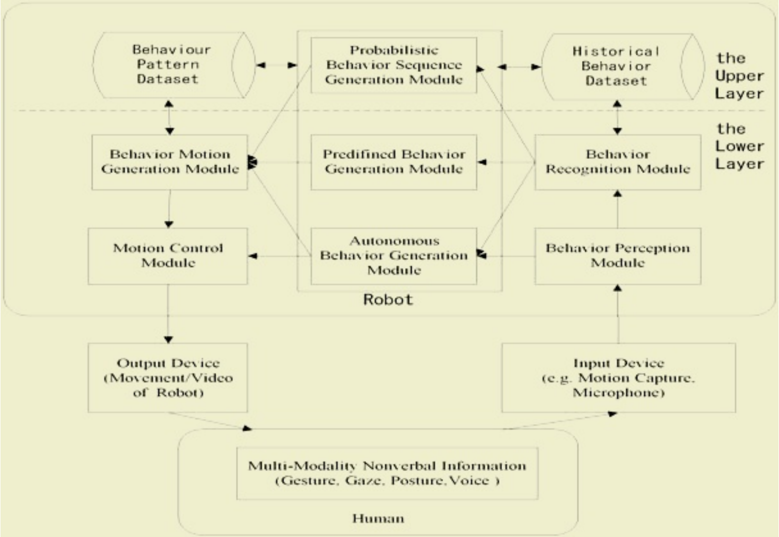

Behaviour Pattern Dataset
Probabilistic Behavior Sequence Generation Module
Historical Behavior Dataset
the Upper Layer
the Lower Layer
Behavior Motion Generation Module
Predifined Behavior Generation Module
Behavior Recognition Module
Motion Control Module
Autonomous Behavior Generation Module
Behavior Perception Module
Robot
Output Device (Movement/Video of Robot)
Input Device (e.g. Motion Capture, Microphone)
Multi-Modality Nonverbal Information (Gesture, Gaze, Posture, Voice)
Human

Fig. 2. Architecture of Robot System

The input of the system is human user's nonverbal behavior such as gesture. The motion capture will be used to input human user's gesture in our experiment. Then the "Behavior Perception Module" preprocesses raw input data, if the data have too much noise to be recognized, the module will transfer control directly to the "Autonomous Behavior Generation Module". This module generates autonomous behavior. "Behavior Recognition Module" recognizes behavior pattern and record past data into historical dataset. "Predefined Behavior Generation Module" generates behavior according to predefined rules. "Probabilistic Behavior Generation Module" generates behavior sequence according to probability values which are calculated from historical behavior dataset. "Behavior Motion Generation Module" generates high level behavior motions, such as nodding, saying "ya", etc. "Motion control module" generates low level motion commands, such as motor commands. Output device will complete robot's movement to make user can see the robot's movements of responses.

3.1 The Upper Layer

The upper layer mainly determines how to transit among different communication situations and how to generate behavior sequences based on a statistical model, i.e. Bayesian Network. [2]

As mentioned in section 2.5, the robot can give proper response to human user by transiting among different communication situation. Each communication situation progresses according to schema based interaction [3]. Respective schema is defined according to human's behaviors following social habits, where the robot acts by computing probability of schema which combines human and robot's behaviors, such as nodding, waving hands, greeting after approaching, gazing at each other, etc.

3.2 The Lower Layer

The lower layer is responsible for establishing robust event recognition, by focusing on rhythmic and repetitive gestures. Since it is relatively easy to extract periodic behavior even from noisy data, this approach may bring about a robust human-robot interaction.

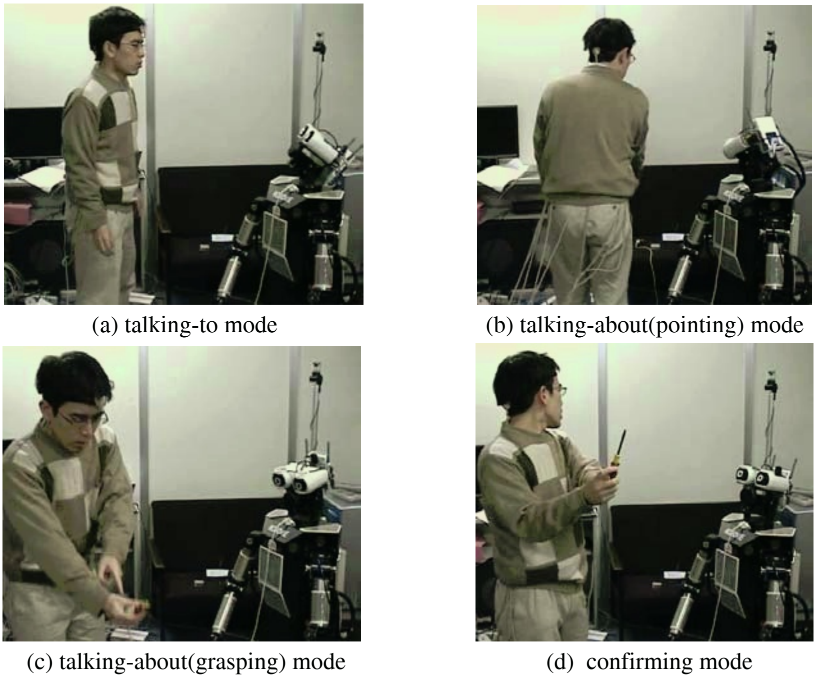
We use the entrainment mechanism [4] to couple human-robot communication. Here entrainment is defined as a kind of phenomenon that the orbits of human user's gestures and the movement of the robot become synchronized during the human-robot interaction.

In the lower layer, recognizing user's attention behaviors play an important role. The attention behavior not only helps to distinguish different communication situations, and transit among different communication situations, but also helps to convey user's intention to the robot. The redundancy of attention behavior is very crucial for recognizing human's intentions.[5] It is because the behavior represented by only one modality is not robust enough. So two kinds of redundancy: redundancy of modality, such as gesture, posture, voice and eye gaze, and redundancy of time, such as repetitive or persistent behaviors, will be considered.

4 Implementation

A waiter robot system developed by Hatakeyama [2] is implemented based on interaction schema, which can establish communication atmosphere before the human-robot communication begins. The human-robot distance, gaze direction, ACK(acknowledgement) action and history of behaviors are processed to generate or modify interaction schemata, then the system determines robot's reaction based on probabilistic reasoning by Bayesian network. However the waiter robot system focus on opening communication situation, it may limit the robot's possible reactions.

A part of our ongoing work is to construct a reporter robot system basing on the listener robot system which is developed by Ogasawara [6]. The listener robot mainly works in passive communication situation and can respond to human speaker by nodding and saying "Ya" when the latter is explaining something to it. The listener robot consists of three main modules: the Attention Behavior Recognition Module, the Communication Mode Estimation Module, and the Robot Behavior Generation Module. The first module recognizes speakers' body motion and speech. It can identify the type of speaker's behavior and object(s) that the speaker pays attention to. In addition, the robot can estimate the intensity of intention based on the redundancy of behaviors by the second module. Using these information, the last module will select most confident behavior according to speaker's mode. Four communication modes between speaker and robot are used: talking-to mode, talking-about mode, confirming mode and busy mode. The robot system implemented with four modes is shown in Figure 3.



(a) talking-to mode

(b) talking-about(pointing) mode

(c) talking-about(grasping) mode

(d) confirming mode

Fig. 3. Communication Modes Used by Listener Robot

Since the listener robot mainly works in passive communication situation, it may not adapt to user actively. The reporter robot will enable system to transit into active situation if necessary. It can recognize which communication situation it is engaged in, and acts according to the situation and human gestures. As its input, initial or previous situation can be determined from the movements of the user and the robot. As the output, the reporter robot generates behavior sequences and response movements according to the next situation. The function of recognizing user's simple motion, such as repetitive gesture, will be performed by lower layer. While complicate behavior sequences will be generated by upper layer.

5 Conclusion and Discussion

In this paper, we addressed the issue of establishing a natural communication environment between a human user and a robot. We proposed a two-layered architecture for building our robot system. We describe how to build the robot that can transit among communication situations and adapt to a human user's action. However since the robot system has not been implemented, we plan to build the robot system and perform more experiments to evaluate the effectiveness of our system.

Since the number of communication situation mentioned in this paper may not enough, it limits available possibility for the robot to generate proper response to the human user. We plan to define more new communication situations to improve the system. What is more, it is insufficient for the robot to adapt to different user. It should be useful to introduce more adaptive ability for the robot.

Generally speaking, we plan to give further research about more communication situation, motion recognition and adaptive ability of the robot in near future.

References

1. Sinder C., Lee C., Kidd C., Engagement During Dialogues with Robots, In Proc. of Symposium on Informatics for Supporting Social Intelligence and Interaction, SSAISB2005 Convention, 27-31, Hatfield, UK (2005)
2. HATAKEYAMA M., Human-Robot Interaction based on Interaction Schema (in Japanese), Master Thesis, University of Tokyo, Japan (2004)
3. HATAKEYAMA M., HATADA T., NISHIDA T., Robot and Human Communication Using Bodily Expressions, SICE Annual Conference 2003, TAI-11-1, Japan (2003)
4. TAJIMA T., Human-Agent Communication of Tacit Intention by Symbol Mapping and Entrainment (in Japanese), Master's thesis, Graduate School of Information Science and Technology, The University of Tokyo, Japan (2004)
5. OKAMOTO M., NAKANO I.Y., and NISHIDA T., Toward Enhancing User Involvement Via Empathy Channel in Human-Computer Interface Design, In Proc. of IMTCI, Poland, (2004)
6. OGASAWARA Y., A Listener Robot Capable of Establishing Joint Attention Based on the Redundancy of Nonverbal Behaviors (in Japanese), Master's thesis, Graduate School of Information Science and Technology, The University of Tokyo, Japan (2005)

Analysis of Conversation Quanta for Conversational Knowledge Circulation

Ken Saito, Hidekazu Kubota, Yasuyuki Sumi, and Toyoaki Nishida

Dept. of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
saitoh@ii.ist.i.kyoto-u.ac.jp

Abstract. In this paper, we present a computational approach to understanding and augmenting the conversational knowledge process. We introduce the concept of the conversation quantization, a technique of approximating a continuous flow of conversation by a series of conversation quanta that represent points of the discourse. To investigate what the nature of conversation quanta is, we attempt to extract conversation quanta from two types of the meeting videos by hand. As a result, we have obtained some profitable suggestions about conversation quanta.

1 Introduction

The conversation is our primary method to communicate with each other. A lot of useful knowledge occurs in the conversation in the real world, however it almost fades away without the support of intelligent media technologies. In this paper, we present a computational approach to understand and augment the conversational knowledge process that is a collective activity for knowledge creation, management, and application where conversational communications are used as a primary means of interaction among participating agents. The key idea is conversation quantization, a technique of approximating a continuous flow of conversation by a series of conversation quanta that represent points of the discourse.

Previous works about conversation quantization surveyed by Nishida [1] include expansions of conversation quanta in knowledge circulation, embodied conversational agents, a virtual immersive environment, spatial content management and automated conversation capture. The former studies contribute a lot to knowledge circulation by using handcrafted conversation quanta. We are here concerned with the nature of conversation quanta from real world conversation on assumption that conversation quanta are utilized for a conversational agent that can talk on behalf of an actual person.

This paper describes the conceptual framework of conversation quantization and two analyses of conversation quanta. We have experimentally extracted conversation quanta by hand from practical situations to investigate their nature and reusability.

2 Conversation Quantization

Conversation quantization is a technique of articulating a continuous flow of conversation by a series of objects called conversation quanta each of which represents a

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 296–302, 2005.

point of the discourse. We define a conversation quantum to be an entity that contains a minimal amount of contextual information. In other words, each conversation quantum makes a minimal sense even though it may be presented in an inappropriate context. The granularity and size of conversation quanta essentially depend on the context and background knowledge of the observer.

Conceptually, the framework of conversation quantization consists of extraction, accumulation, processing, and application of conversation quanta (Fig.1). The extraction of conversation quanta results from identification and encoding of coherent segments of interactions in a conversational situation. The extracted conversation quanta are accumulated in a server, processed whenever necessary, and applied to other conversational situations. The application of conversation quanta involves information retrieving, knowledge sharing and hands-on learning. In these situations, an embodied conversational agent is a good mediator because of its communicative abilities.

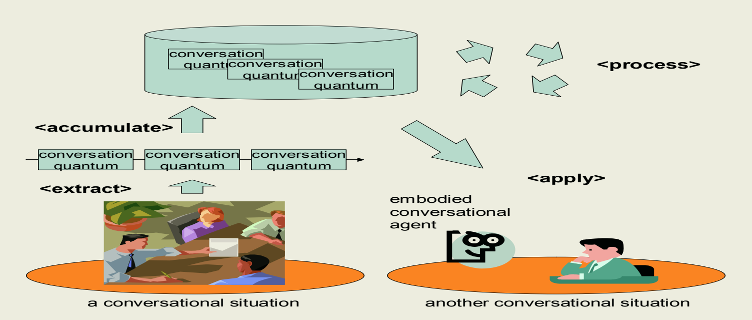


Fig. 1. The concept of conversation quantization

In this paper, we investigate the nature of the conversation quanta focusing on the virtualized-ego agent. The virtualized-ego is a conversable agent that functions as an alter-ego. A user can talk with a virtualized-ego of an actual person whenever and wherever he likes [2]. A virtualized-ego is not a virtual character but a virtualized human that has autobiographical memory of an actual person. It is important that our aim is creating a conversational agent that can talk about informal knowledge something like personal experience. The virtualized-ego is expected to decrease the time-related, spatial and social constraint of the conversation.

3 Nature of Conversation Quanta

The implementation of conversation quantization depends on the data structure for representing conversation quanta. One could use plain video clips as representation, but the efficiency in retrieving and processing would be quite limited and a large cost would be require for retrieving, editing, and applying the conversation quanta. Alternatively, a deep semantic representation using logical formulas or case frames would not be ideal due to the expense and the limited capability of representing nonverbal information. A reasonable implementation appears to be to use annotated videos and images to represent a conversation quantum.

Firstly, we targeted a conversation using slides. We simulated conversation quantization by hand to investigate the nature of conversation quanta in a real situation. We gave shape to the concept of conversation quanta as follows:

1. Setting up a practical conversational situation
2. Capturing conversation by video camera
3. Extracting conversation quanta from the video stream by hand

We obtained the two types of video. One is the video in which two members of authors participated (Video A). The other is the video in which 4 people (including one of authors) participated (Video B). In the following, we will empirically analyze the nature of conversation quanta by extracting the conversation quanta from these videos, and creating new conversational contents using these conversation quanta.

3.1 ANALYSIS 1 (Video A)

Video A consists of 3 meetings between two of the authors, a master course student (subject A) and a postdoctoral fellow (subject B). Each meeting was held in a different place and at different times. Each of them talked using PowerPoint slides on mobile PC (with a web camera and a microphone) to capture his voices, faces and context. As a result of these meetings, we obtain three and a half hours video of subject A and subject B. In their conversation, presentational style and discussion style are half-and-half, and their topics were about conversation quantization – its history, problems, approaches, systems and so on.

Here, we suppose that we can create virtualized-ego by quantizing the video and mixing the quanta. To confirm this supposition, we made an experiment to extract conversation quanta arbitrarily from these videos and create a new presentation video in which the virtualized-ego of the participants talk about their study.

We proposed the first approximate model of extracting conversation quanta (Fig.2). We divided videos by the following policy:

1. Divide at the point of the transition of slides (Fig.2 Division 1)
2. Divide at the point of a start of dialogue (Fig.2 Division 2)

Firstly, the video is divided at the point of the transition of slides (Fig.2 Division 1) because our attempt suggested that speeches are almost coherent in a slide. The second division point (Fig.2 Division 2) is a start of dialogue. We suppose that a video clip from the start of dialogue to the end of a slide is a conversation quantum that is dialogue style. A conversation quantum is stored into the archive of a speaker as a general rule. Only when a quantum is a dialogue style, it is stored into the archive of every speaker.

Using this model, we extracted conversation quanta from Video A. We have got 41 quanta for subject A and 66 quanta for subject B. Table 1 shows the number of conversation quanta in the archive of each subject. “Single speech” means the quanta include only the subject, and “Dialogue” means the quanta include two subjects. The number of dialogue quanta of subject A is same as that of subject B of course.

We have simulated a virtualized-ego system based on these conversation quanta. We arranged conversation quanta of subject A on the assumption that the system talks with a user on behalf of subject A. Fig.3 shows the overview of our simulation. Firstly, a user comes in front of a system screen where the face of subject A is displayed. Here, the system begins to talk on behalf of subject A when the user asks for his interest ("Greeting"). The system talks by arranging past conversation quanta that are related to the interest of the user ("Quantum 1" and "Quantum2"). While the system is talking, the user can ask any questions ("Question"). Then the system can answer the question by searching an answering conversation quantum ("Quantum3"), and keep on talking (Quantum4).

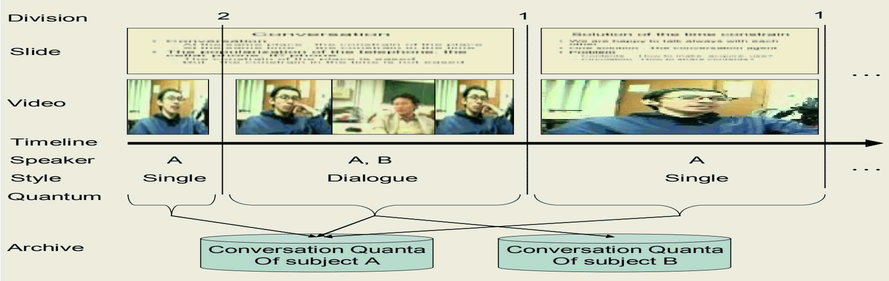


Fig. 2. The first approximate model for extracting conversation quanta

Table 1. Conversation quanta from three and a half hours videos

Subject	Single speech (total time)	Dialogue (total time)
A	24 quanta(16 minutes)	17 quanta(21 minutes)
B	49 quanta(35 minutes)	17 quanta(21 minutes)

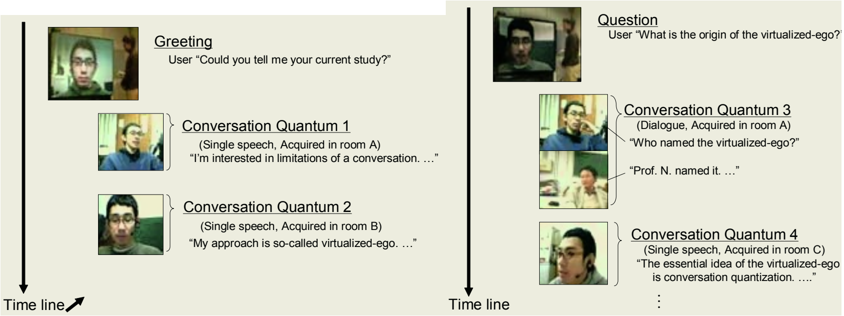

Greeting
User "Could you tell me your current study?"
Conversation Quantum 1
(Single speech, Acquired in room A)
"I'm interested in limitations of a conversation. …"
Conversation Quantum 2
(Single speech, Acquired in room B)
"My approach is so-called virtualized-ego. …"
Time line
Question
User "What is the origin of the virtualized-ego?"
Conversation Quantum 3
(Dialogue, Acquired in room A)
"Who named the virtualized-ego?"
"Prof. N. named it. …"
Conversation Quantum 4
(Single speech, Acquired in room C)
"The essential idea of the virtualized-ego is conversation quantization. …."
Time line

Fig. 3. A simulation of a conversational video content using conversation quanta

We have obtained some profitable suggestions about conversation quanta from the simulation and analysis above. First, conversation quanta which depended on context are reusable in the situation where a user is familiar with the original situations of the conversation quanta. In Fig.3, Quantum1, Quantum2, Quantum4 were acquired in different rooms. Thus, we can make new conversational content from the past conversation quanta that were got in different situations. When we searched conversation quanta that are suitable for a user, thinking about the background knowledge of the user was very important. The conversation in Fig.3 left fragment, however, it could be complemented by the user because he is a colleague of the speakers on the screen. Second, a dialogue style quantum which contains speedy Q&A pair, jokes is interesting. They have good points of conversation such as conversational rhythms, presence, and dynamics. In addition, on the viewpoint of virtualized-ego, a quantum which contains individual experiences and know-how is interesting too.

3.2 ANALYSIS 2 (Video B)

One of authors had the meeting in which four people participated (subject C, D, E, F). In this meeting, a doctor's course student of our laboratory (subject C) presented his studies with PowerPoint slides. This meeting was captured with a digital video, and then we have got one and a quarter hours video of the meeting. In their conversation, ratio of presentational style to discussion style is about 1:5.

In addition to the first approximate model, we adopted the third division point for dividing detailedly. The third division point (Fig.4 Division 3) is the part in which no one speaks. In other words, this is silence from the end to the start of speech. We have proposed the second approximate model of extracting conversation quanta (Fig.4). Namely we divided videos by the following policy:

1. Divide at the point of the transition of slides (Fig.4 Division 1)
2. Divide at the point of a start of dialogue (Fig.4 Division 2)
3. Divide at the part in which no one speaks (Fig.4 Division 3)

On the third division point, we divided the part in which silence is more than 4 seconds. Table 2 shows the number of conversation quanta and quanta's average time.

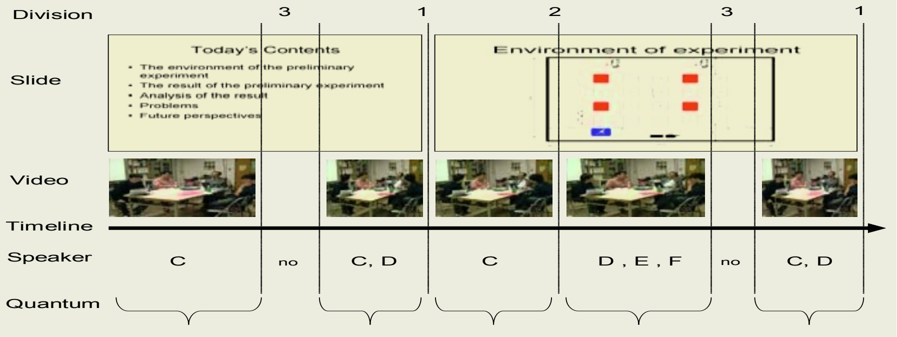


Fig. 4. The second approximate model for extracting conversation quanta

Table 2. Conversation quanta from Video B

Threshold of silence	Division	Silence (average time)	Quantum (average time)
More than 4 sec	111	47 (6.9 sec)	65(66 sec)

As a result of the analysis, we found that the Video B includes two types of conversation quanta. One is the presentation in which subject C explained his slide. The other is the discussion by all subjects.

The presentation style quantum arises almost at the head of a slide at 71 %(11/14). The presentation style quantum does not very depend on the context in compare with the discussion style quantum. The presentation quantum gives the context to the discussion style quantum in right after because the presentation shows what to discuss. Some discussion style quanta require the presentation style quantum to be understood.

Although it is difficult to understand the discussion style quantum when a user isn't given context, it is important and interesting. Because the discussion style quantum contains individual experience and know-how and joke.

4 Discussion

In the previous sections, we have discussed the nature of the conversation quanta on the viewpoint of virtualized-ego. As a result of two analyses, we obtain some suggestions about conversation quanta.

The conversation quanta which contain following contents, speedy Q&A pair, jokes, and individual experiences and know-how, are important and interesting in reusing conversation quanta. The virtualized-ego which is generated from conversation quanta that depend on context would be understandable for the user who shares the context. For example, a community member could easily complement fragmentized conversation of the members in the same community. This supposition was confirmed by Hirata [3] in only text conversation fragments, by Kubota [2] in only voice conversation fragments. In this paper, we obtain the result above mentioned about video conversation fragments.

There are interesting works about conversation quantization. Conversation quanta can be extracted from the real world conversation by expanding the ubiquitous sensor room that is proposed by Sumi et al [4]. For spatio-temporal management of conversation quanta, Kubota is developing a system called the Sustainable Knowledge Globe [5]. Virtualized-ego [6] that can talk on behalf of an actual person (as mentioned in Section 3.1) is a good utilization of conversation quanta. Video and sound collage system of one's experience [4] would also be another good application.

The study of topic extraction from conversation has been growing. A deep understanding of discourse structure is indispensable to extract essence of conversation automatically. Shibata et al. [7] study the discourse analysis of the cooking program video by using linguistic and visual information. Our research object doesn't like the cooking program which is controlled by a video director but the casual conversation in any situations, so it is very difficult to understand the discourse structure automatically, especially the correspondence structure when there is the omission in the indication word in the video. We aim to make the conversation quantization feasible by supporting humans to understand topics in video in the loop of the conversation quantization (extraction, accumulate, apply, process) such as Q&A system of Ego Chat [6].

There are many interesting work left for the future research. Among others, we need to build a more detailed and elegant theory of conversation quantization. A more sophisticated theory of conversation quanta will permit us to better design the representation and basic operation for conversation quanta. It may well enable us to predict the cost and effect of building a conversation system based on conversation quantization.

5 Conclusion

In this paper, we have presented a computational approach to understanding and augmenting the conversational knowledge process. We have introduced the notion of conversation quantization, a technique of approximating a continuous flow of conversation by a series of conversation quanta that represent points of the discourse. We obtained profitable suggestion about the nature of conversation quanta by extracting them from practical situations by hand.

References

1. Toyoaki Nishida: Conversation Quantization for Conversational Knowledge Process, Special Invited Talk, S. Bhalla (Ed.): DNIS 2005, LNCS 3433, Springer, pp. 15-33, 2005.
2. Hidekazu Kubota, Toyoaki Nishida, Tomoko Koda: Exchanging Tacit Community Knowledge by Talking-virtualized-egos. Fourth International Conference on AUTONOMOUS AGENTS (Agents 2000, Barcelona, Catalonia, Spain. June3 -June 7), pp.285-292, 2000
3. Takashi Hirata, Hidekazu Kubota, and Toyoaki Nishida: Talking virtualized egos for dynamic knowledge interaction. In Toyoaki Nishida, editor, Dynamic Knowledge Interaction, chapter 6, pages 183-222. CRC press, 2000
4. Y. Sumi, K. Mase, C. Mueller, S. Iwasawa, S.Ito, M. Takahashi, K.Kumagai, Y.Otaka: Collage of video and Sound for Raising the Awareness of Situated Conversations. In Proceedings of International Workshop on Intelligent Media Technology for Communicative Intelligence (IMTCI2004) (2004) pp.167-172
5. Kubota, H., Sumi, Y., Nishida, T.: Sustainable Knowledge Globe: A System for Supporting Content-oriented Conversation, in Proceedings of AISB 2005 Symposium Conversational Informatics for Supporting Social Intelligence & Interaction, pp.80-86, 2005
6. Hidekazu Kubota, Jaewon Hur, Toyoaki Nishida: Agent-based Content Management System. In Proceedings of the 3rd Workshop on Social Intelligence Design (SID 2004), CTIT Proceedings (2004) pp.77-84
7. Tomohide Shibata, Masato Tachiki, Daisuke Kawahara, Masashi Okamoto, Sadao Kurohashi, and Toyoaki Nishida: Structural Analysis of Instruction Utterances using Linguistic and Visual Information, In Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES2004), Wellington, New Zealand (2004.9) pp. 393-400

An Intelligent Approach of Packet Marking at Edge Router for IP Traceback*

Dae Sun Kim[1], Choong Seon Hong[2,**], and Yu Xiang[3]

School of Electronics and Information, Kyung Hee Univerity, 449-701, Korea
{dskim,cshong}@khu.ac.kr, skipperyu@hotmail.com

Abstract. With the help of real source identity in packets, network security system can intelligently protect and counteract the attacks. Packet marking is an important method of source identification, and there are some issues on it. For large amount of packets, analysis time and complicated computation are necessary while detect marking information. This paper focuses on this direction, and proposes a simple and efficient method to mark all packets belonging to upstream traffic with a deterministic, plain form identity. With this approach, we just need low processing power on some specific edge routers as well as a little extra network traffic to settle it. Furthermore, distilling mark from packets is easy since the mark is in plain text format.

1 Introduction

Attackers routinely disguised their location using incorrectly, or spoofed IP source address. A great amount of effort has been made upon traceback to get the source information. Ideally, full-path traceback is a good way. But according to [1], packets may be delivered along different path to the destination (load balancing or unwanted isolation of the network routing) sometimes, only the ingress interface on the router closest to the source is almost same. The authors of [5] divides traceback technologies into two categories according to tracing clues: Traceback across stepping-stones and IP traceback. The first type is mainly for connection trace. For against DoS or DDoS, IP traceback is useful. It focuses on packet trace: Logging, ICMP Trace, probabilistic packet marking (PPM), Algebraic approach, Tunnel technologies, etc. Actually there are still some limitations in these approaches: large amounts of packets and complicated computation are necessary. Convergence procedure is relatively slow and background noise and spoofing marked packets may affect their performance.

2 Related Work

Source Path Isolation Engine (SPIE) [8,9] uses hash-based technique to gain every packet's information for IP traceback that generates and stores audit trails

* This work was supported by University ITRC Project of MIC

** Dr. C.S. Hong is the corresponding author

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 303–309, 2005.

for inquery, and can trace the origin of a single IP packet delivered by the network in the recent past. About 0.5 percent of the link capacity per unit time in storage is needed. For realization all routers need to be controlled by specific manager. Besides of the corporation among all ISPs, wide deployment in whole Internet also is a big challenge. In Pi [6] is also a per-packet deterministic mechanism that allows the victim side to filter out packets matching the attacker's identifier embedded in each packet. It uses the digest of path and routers IP address as the identity. Each packet traveling along the same path carries the same identifier. It uses 16 bits of Identity field of IP header to store mark. Since the space is not enough for mark, tradeoff has to be adopted upon some hands, such as efficency and reliability. Enough quantity routers along all possilbe paths are needed. The authors had given a experience to show that the scheme is available in case of half of all related routers along attacking path. IP Traceback-based Intelligent Packet Filtering [7] is an integrated infrasture assembled with some approaches. And it is based on focused on filterring out the majority of DDoS traffic to improve the overall throughput of the legitimate traffic. With the help of PPM, the victim side can find the attacking paths and then filters out these "infected" path traffic in some degree. But this scheme needs rigorous conditions. Its EPM function modle must be deployed almost on all source side routers as well as all routers on the defence line for victim network are with PPF function. This prefigures the cost is very high. Deterministic Packet Marking (DPM) [1] is a novel packet-marking algorithm with all packets marked at edge routers. Like some approachs, the 16-bit ID field and the reserved 1-bit Flag in the IP header are used to store mark. The biggest difference is that incoming interface's IP address (32 bits) will be stored in two packets. The coding in the ID Field assumes that there are almost no IP fragments in the Internet supported by empirical traffic analysis less than 0.5 percent of all packets in Internet are fragmented in [4]. But SPIE sends mark by extra network bandwidth overhead for every traffic packet. Pi marks every packet with path information for reconstructing attack path map without extra network overhead. [7] inserts path identity into packet as mark in probabilistic. DPM marks every packet at source side with plain text of IP address. We found that the scanty space for storing mark in packet is a key reason among almost all approaches. Our scheme is inspired from the above works and is absorbed in finding more available space for storing mark.

3 Overview of Proposed Packet Marking Scheme

After the literature survey of popular IP traceback techniques, we claim that complex and more restriction traceback will suffer from scalability and deployment problem. An optimal and simple scheme should be introduced with lower processing and network traffic overhead. Our proposal draws inspirations from DPM scheme and also marks all upstream packets at one of source side routers belonging to edge router of subnet or domain as shown in Figure 1. If we allocate the subnet a global identity (GID) number as mark directly and store the mark in all output traffic packets, then destination receiving these packets can detect the source easily. However, storing a global identity number need more

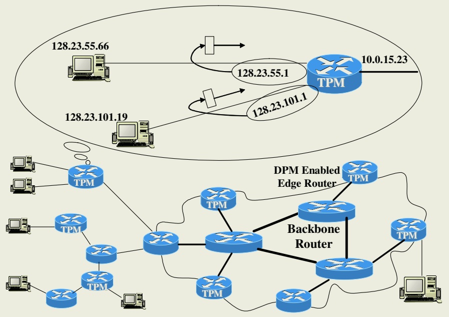


Fig. 1. TPM marks subnet identification at incoming interface of router which is connected to gateway of this subnet into all upstream traffic packets

space than 16 bits, it seem impossible that satisfy this asking from IP header fixed part directly. This is the key reason why most of current approaches use compression pattern for mark. In general viewpoint, there is no more space in IP header enough to store a global identity. Fragmentation related fields may be the last hope. Identity field is just only used for the fragmented packet, and Offset field is the same here with a little difference. If it is certain that a packet is not a fragment or DF is set, this field can be definitely used. In this case we can put the global identity of subnet in 30 bits space as in Figure 2. In general

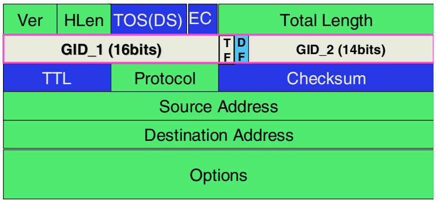


Fig. 2. Storing Mark of Un-fragmented Packet in IP header fixed part

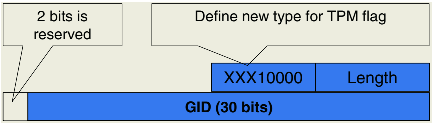
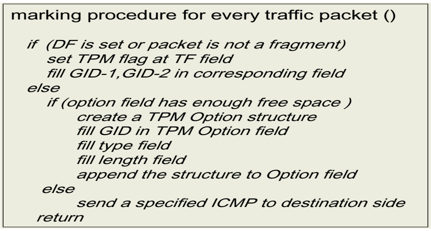


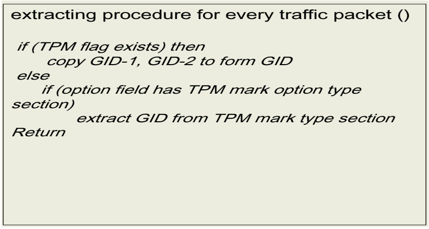
Fig. 3. Fragmented Packet Mark in Option Field

direct uses of IP address as global identity is the best and ideal choice but as shown in figure 2, there are only 31 bits that we can use. We can ignore the last one bit even more bits because IP addresses have only difference in last a few bits must locate in the same subnet or domain. The underlying reason we can perform the ignore action is that all we need to distinguish subnets instead of hosts. Total 30 bits resource is big enough to identify all subnets in Internet. Final receiver should know that if the packet is not fragmented. Thus we keep the DF (do not fragmentation) as it old value. The one bit at left side of DF is used for marking flag. For fragmented packet, instead, Option field is used. In this scheme we define a new sub-section in Option field with a new flag for IP traceback as shown in figure 3. Total overhead for the new sub-section is six bytes. The marks for both un-fragmented case and fragmented case are in plain text form. Very easily works are needed both on marking at source side and extracting mark at destination side. The mark function is very easy to implement on routers. One of the possible algorithms of marking and extracting function looks as figure 4 and figure5.



```
marking procedure for every traffic packet ()

  if  (DF is set or packet is not a fragment)
      set TPM flag at TF field
      fill GID-1,GID-2 in corresponding field
  else
      if (option field has enough free space )
            create a TPM Option structure
            fill GID in TPM Option field
            fill type field
            fill length field
            append the structure to Option field
     else
            send a specified ICMP to destination side
    return
```

Fig. 4. Make Inserting Algorithm



```
extracting procedure for every traffic packet ()

 if (TPM flag exists) then
        copy GID-1, GID-2 to form GID
 else
       if (option field has TPM mark option type
section)
             extract GID from TPM mark type section
Return
```

Fig. 5. Make Extracting Algorithm

4 Discussion and Analysis

As well-known, the usage of fragmentation is decided by MTU size [3]. Fragmentation will degrade the efficiency and performance of Internet. In order to improve the performance of whole network, RFC1191 specifies a path MTU discovery protocol. At present this protocol is widely using in Internet, so majority of traffic do not experience fragmentation and normally DF is set. But there still has exception. A few IP protocol stack did not support this protocol. For instance, an un-fragmented packet (its MTU size is X, also it was not compliant PMTU) travels through the edge ingress router into backbone network and then pass through the downstream router. At this time if a more low speed data link media exists at the side of destination and its MTU is smaller than X, then fragmentation appears. It will destroy mark even the traffic has not been disrupted. Since the protected servers are valuable and important servers which access pattern is impossible lower than Ethernet link, thus the paths from it to the edge routers are not low bandwidth transmission media predicatively. According to the traffic analysis of Internet, nearly all packets size is not bigger 1500 bytes. In a word, in the paths from protected servers to marking point, the fragmentation case does not exist even in very special situation. As shown in figure 1, in TPM scheme it is unnecessary to add TPM function to all source side routers. Instead, just ensures all traffic packets to be marked one times and just one times. At the same time, instead of marking at edge router of backbone network, chooses TPM marking point close to hosts as possible. This rule is easy to be satisfied obviously. So that the granularity of source information can support the traffic flow restriction while against DoS or DDoS attacks.

5 Simulation Results and Evaluation

In this section we will demonstrate the simulation of our scheme and the comparison with DPM scheme. Based on OPNET 8.0.c remote server version under

MS Windows2000 professional version and Microsoft Visual C++ 6.0 compiling environment, The first simulation tests the impact on introducing the TPM function to existing routers. In our simulation, Database Access, email, http, ftp traffic are mixed and router's CPU background utilization is set to 30 percent for simulating the Internet traffic as possible. As figure 6 shown in TPM case is a very little higher than result points of without TPM case in Y-axis. Although smaller scale has been used in Y-axis, the difference between the two cases is too small to distinguish them easily. This means that the CPU utilization of edge router with TPM function embedded is very little higher than the CPU utilization of edge router without TPM function. The result in figure 7 shows nearly the same as figure 6. In second phase we compare our scheme with DPM scheme. Since DPM scheme must consider attacks might alternate the source IP address, we include this factor by adding a little part (0.8 percent)of source IP address changed. From figure 8 shown, TPM approach gains the average detection speed around 13 packets per seconds. We use two independent subnets as traffic source. Each subnet continually generates specific packets in uniform distribution pattern with interval time 0.1 to 0.2. That is to say TPM approach has almost detected all marks of received packets. On the contrary, the result line of DPM approach is around 6 marks per seconds in mark detection rate.

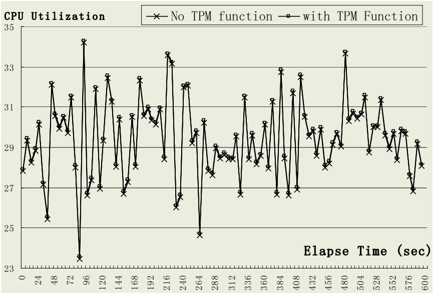


Fig. 6. CPU Utilization Comparison

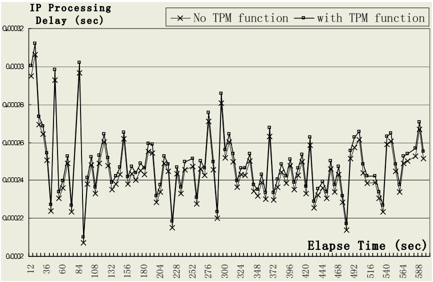


Fig. 7. IP Processing Delay Comparisons

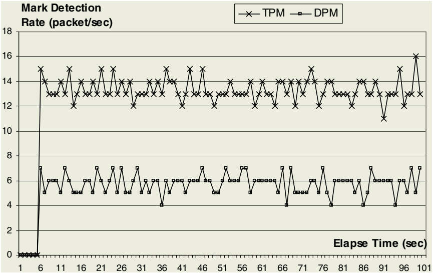


Fig. 8. Mark Detection Speed Comparisons

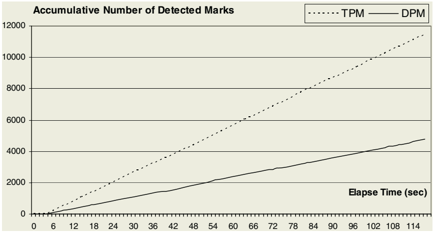


Fig. 9. Accumulative Number of Detected Packets

This is lower than half of TPM mark detection rate. There are several reasons for this difference. At first, DPM mark detection rate at most equal the half level as TPM approach even under ideal pre-arranged packets arriving sequence. Because DPM approach at least need two different mark parts to assemble a complete mark. Next is that the packets arrived in random. If the two or more packets carry with the same half of a complete packet continually, the detection rate must be decreased certainly. Another reason is that some packets with frequent change source IP address do not contribute mark detection as we discussed in previous part of this section.

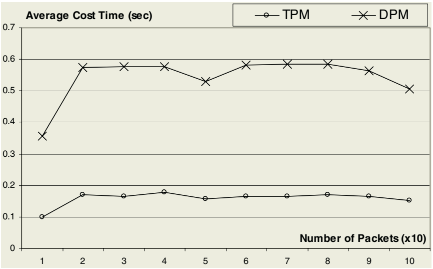


Fig. 10. Average Time Cost for Extracting Mark vs. Arriving Packets Speed around 5 to 10 pps

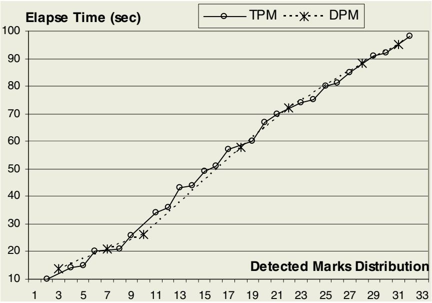


Fig. 11. Time Distribution of Detected Marks Comparison vs. Arriving Packets Speed around 0.5 to 1 pps

Therefore in figure 8 parts of marked packets could not contribute to assemble right mark. Furthermore instable source IP address will embarrass the mark's availability. It leads invalid mark packets even more than mark detected rate. figure 9 is about shows intuitionistic comparison between DPM scheme and our scheme upon number of detection rate by simulation procedure. Figure 10 is about the spending or cost of time for gaining the right mark. The last simulation item intends to compare the number of extracted marks and extracted time. The arriving packets number is reduced in order to convenient observing. Look at the result in figure 11, Some packets with mark were not detected in DPM scheme. Also the number of detected mark is smaller than half of TPM scheme detection. This confirms that during the same time TPM scheme distilled more marks than DPM scheme obviously.

6 Conclusion and Future works

In this paper we have introduced a new approach to IP traceback. This approach effectively addresses shortcoming of existing techniques in some degree. The important contributions of TPM lie in plain text form and a complete mark existing in every packet. It is not necessary computation to generate a different mark in real time for each wanted marking packet as current most approaches. This is big difference comparing with many packet marking approaches which generate digest information as mark using Message Authentication Code algorithm before each marking action. Marking all traffic packets can obtain the fastest speed of detection source information at destination sides. This approach is also efficient against various types of attacks, not only for DoS or DDoS attacks since all packets have trustworthy source side information, furthermore mark in plain text can be directly read and processed by security systems easily. We plan to do more deep investigation and experiments related to Internet traffic analysis, such as MTU size distribution, Option field using situation in Internet roundly to gain more supporting data for TPM scheme.

References

1. A. Belenky and N. Ansari, “IP Traceback with Deterministic Packet Marking,” IEEE Communications Letters, Vol. 7, No. 4, pp. 162-164, April 2003
2. S. Savage, D. Wetherall, A. Karlin, and T. Anderson. “Network support for IP traceback” IEEE/ACM Trans. Networking, vol. 9, pp. 226-237, June 2001.
3. Behrouz A.Forouzan, Sohia Chung Fegan, “TCP/IP Protocol Suite” Chapter 7, pp 152-153
4. Stefan Savage, David Wetherall, Anna Karlin and Tom Anderson, “Practical Network Support for IP Traceback”, Proc. of the ACM SIGCOMM conference, August 2000, Stockholm, Sweden, Computer Communication Review Vol. 30, No 4, October2000
5. “A Little Background on Trace Back”.
URL:discovery.csc.ncsu.edu/ pning/Courses/ csc774/on-trace-back.pdf
6. Abraham Yaar, Adrian Perrig and Dawn Song, “Pi: A Path Identification Mechanism to Defend against DDoS Attacks”, In Proceedings of the 2003 Security and Privacy Symposium, May. 2003
7. Sung, M. and Xu, J. “IP Traceback-based Intelligent Packet Filtering: A Novel Technique for Defending Against Internet DDoS Attacks”, IEEE Transactions on Parallel and Distributed Systems, vol 14, no 9, pp. 861–872, September 2003
8. Alex C. Snoeren et al., “Hash-Based IP Traceback”, Proc. of the ACM SIGCOMM conference 2001, San Diego, CA, Computer Communication Review Vol. 31, No 4, October 2001, pp. 3-14
9. A. Snoeren, C. Partridge, L. Sanchez, C. Jones, F. Tchakountio, B. Schwartz, S. Kent, and W. Strayer. Single-Packet IP Traceback. In ACM/IEEE Transactions on Networking, vol. 10, no. 6, December 2002.

A Covert Timing Channel-Free Optimistic Concurrency Control Scheme for Multilevel Secure Database Management Systems

Sukhoon Kang[1] and Yong-Rak Choi[2]

[1] Department of Computer Engineering, Daejeon University
96-3 Yongun-Dong, Dong-Gu, Daejeon, Korea 300-716
shkang@dju.ac.kr

[2] Department of Computer Engineering, Daejeon University
96-3 Yongun-Dong, Dong-Gu, Daejeon, Korea 300-716
yrchoi@dju.ac.kr

Abstract. This paper presents a set of multilevel-secure optimistic concurrency control (*MLS/OCC*) scheme that has several desirable properties: If lower-level transactions were somehow allowed to continue with its execution in spite of the conflict of high-level transactions, covert timing channel-freeness would be satisfied. This sort of optimistic approach for conflict insensitiveness and the properties of non-blocking and deadlock freedom make the optimistic concurrency control scheme especially attractive to multilevel-secure transaction processing. Unlike pessimistic approaches, the *MLS/OCC* scheme never delays or rejects an operation submitted by a lower-level transaction which is passed the mandatory access control. Instead, the read and write operations are processed freely without updating the actual database. Therefore, it is reasonable to assert that *MLS/OCC* scheme is allowed to avoid the abort of lower-level transactions in order to close covert timing channel, nevertheless guaranteeing conflict-preserving serializability. The basic refinement philosophy for the solution on starvation problem is an incorporation of multiple versions of low-level data into *MLS/OCC*. This kind of intelligent channel-free concurrency control scheme satisfies the B3 or higher level of the US TCSEC requirements.

1 Introduction

A *multiple-level-secure* database management system (MLS/DBMS) is a secure database manager which is shared by users of more than one clearance level and contains data of more than one classification level. An MLS/DBMS is different from a conventional DBMS in two respects: (1) every data item controlled by an MLS/DBMS has associated with it, perhaps indirectly, a unique classification level, and (2) a user's access to data must be controlled on the basis of clearance and classification.

The concurrency control requirements for transaction processing in an MLS/DBMS are different from those in conventional transaction processing systems with respect to inclusion of covert-channel freeness. In particular, there is the need to coordinate transactions at different security levels avoiding both potential *covert timing channels* and the *starvation* of transactions at high security levels. For instance, suppose that a low-level transaction attempts to write a data item that is being read by a higher-level transaction. A covert timing channel arises if the low-level transaction is

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 310–316, 2005.

either *delayed or aborted* by the transaction scheduler. In addition, the high-level transaction may be subject to an indefinite delay if it is forced to abort repeatedly. The user responsible for initiating the aborted transaction must be notified of its unsuccessful termination. But this notification constitutes a flow of information from the DBMS to a low-level user based on the activity of a high-level transaction, and such an information flow may be readily exploited to divulge a sensitive information between conspired transactions.

2 A Covert Timing Channel-Free Optimistic Concurrency Control Scheme for Multilevel Secure Database Systems: *MLS/OCC*

2.1 Applicability of Intelligent Control Scheme for Covert Channel-Freeness

If lower-level transactions were somehow allowed to continue with its execution in spite of the conflict of high-level transactions, *covert timing-channel freeness* would be satisfied. This sort of optimistic approach for conflict insensitiveness is the basic principle behind the set of multilevel-secure optimistic concurrency control (*MLS/OCC*) schemes [1].

An advantage of the optimistic concurrency control (OCC) schemes [2] is their potential to allow a higher level of concurrency. Optimistic concurrency control for multilevel-secure databases (*MLS/OCC*) can be made to work by ensuring that whenever a conflict is detected between a higher-level transaction (T_j) in its validation phase and a lower-level transaction (T_i), T_j is aborted, while T_i is not affected. Ideally, OCC has the properties of non-blocking and deadlock freedom. These properties make the OCC scheme especially attractive to multilevel-secure transaction processing.

However, the original OCC innately possess the clumsy definition of conflict, and thus some transactions can be aborted unnecessarily. This sort of unnecessary abort problem that is caused due to clumsy definition of conflict should be eliminated. Consider a set of transactions that are concurrently executed as shown in Example 1.

Example 1 (Unnecessary Aborts Problem Due to Clumsy Definition of Conflict): The execution of transaction $T_2(S)$ conflicts with $T_1(U)$, since its read-set overlaps with the write-set of $T_1(U)$. Using the validation scheme of original OCC, $T_2(S)$ should be aborted during its validation. However, since $T_2(S)$ reads x after $T_1(U)$ has written it, the serializability between $T_1(U)$ and $T_2(S)$ is guaranteed in the order of $T_1(U) \rightarrow T_2(S)$. Restart of $T_2(S)$ is unnecessary. (See Fig. 1.).

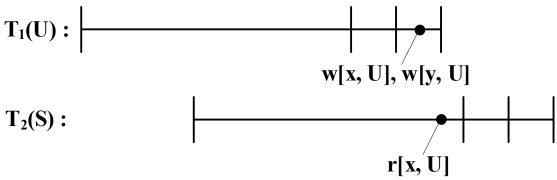


Fig. 1. Execution Schedule for Concurrent Transactions

The basic refinement philosophy for the solution on unnecessary aborts problem is to incorporate the advantage of timestamp ordering into *MLS/OCC* mainly for transaction validation. In this sense, we call this approach *MLS/OTSO.* This variant is shown to preserve the security semantics of the *MLS/OCC* while significantly reducing the cost of its validation phase. Write timestamps can be tested when items are read as well as during validation, thus allowing transactions to detect certain inevitable restarts earlier in their execution.

Note that *MLS/OCC* and *MLS/OTSO*, like all single-version solutions, suffer from starvation due to the fact that if low-level transactions alter a low-level data item sufficiently often then high-level transactions will be unable to successfully read that data item. The basic refinement philosophy for the solution on starvation problem is to incorporate multiple version into *MLS/OTSO*. In this sense, we call this approach *MLS/OMVTO*.

2.2 Covert Channel-Free Validation Phase Algorithm in *MLS/OCC*

Optimistic concurrency control for multilevel-secure databases can be made to work by ensuring that whenever a conflict is detected between a higher-level transaction T_j in its validation phase and a lower-level transaction T_i, the higher-level transaction is aborted, while the lower-level transaction is not affected.

Validation is performed as follows: Let $RC(T_i)$ be the set of recently committed transactions for a transaction T_i, i.e., those transactions which commit between the time when T_i starts executing and the time at T_i enters the critical section for validation. For each transaction T_i the system keeps track of the set of objects read from the database ($RS(T_i)$) and of the set of objects written $WS(T_i)$. T_i is validated if $RS(T_i) \cap WS(T_{RC}) = \varnothing$ for all transactions $T_{RC} \in RC(T_i)$.

If T_i's validation is successful, the transaction is assigned a unique transaction number $TNr(T_i)$. For this purpose a monotonically increasing transaction counter, *TNC*, is maintained. Let TNR_{start} be the highest transaction number at the start of transaction T_j, and let TNR_{finish} be the highest transaction number at the start of validation.

If T_j is validated, its updates are applied to the database; otherwise, it is restarted. Intuitively, T_i is allowed to commit if and only if no other transaction has updated any data items which T_i reads during the time while it was performing its reads and computing its database updates.

Algorithm 1: Covert Channel-Free Validation Phase Algorithm in *MLS/OCC*

```
procedure validate($T$);
begin
    valid:=true;
    for $TNR$ from $TNR_{\text{start+1}}$ to $TNR_{\text{finish}}$ do /* for each $T_{RC} \in RC(T)$ do */
            if $\{RS(T_j) \cap WS(T_i)\} \cup \{RDS(T_j) \cap WS(T_i)\} \neq \varnothing$
             /* for each $x_r \in RS(T)$ do if $x_r \in WS(T_{RC})$ */
            then valid:=false; exit;
    if valid then
            begin
              (write); /* commit $WS(T)$ to database */
```

```
9.              $TNr(T_j)$:=$TNC$;
10.             $TNC$:=$TNC$+1;
11.           end;
12.   else
13.           invoke (read-down) conflict resolution;
14. end;
```

We note that *readset* and *writeset* of a transaction T_i are defined as $RS(T_i)$ and $WS(T_i)$ respectively. We also define read-down set of transaction T_j as the set $RDS(T_j)$ of all T_j's read-down data items and we also define the set of all read-down data items X_{ji} that T_j reads down from transaction T_j. Then we have the following condition: $X_{ji}=RDS(T_j) \cap WS(T_i)$ or $\varnothing$. As a consequence, the equal security level test $\{RS(T_j) \cap WS(T_i)\}$ includes the different security level test $\{RDS(T_j) \cap WS(T_i)\}$.

2.3 Algorithmic Complexity

Let *degree of concurrency* (DC) be the (TNR_{finish} - TNR_{start}), where TNR_{finish} is the highest transaction number at the start of validation and TNR_{start} is the highest transaction number of transaction T_j. Let *transaction size* (TS) be the sum of readset of transaction, T_j, and writeset of transaction, T_i, i.e., TS = $|RS(T_j)|$ + $|WS(T_i)|$. Algorithm 1 iterates |DC| times for each transaction T_{RC}. Through the iteration for each T_{RC}, Algorithm 1 searches T_{RC} in $\{RS(T_j) \cap WS(T_i)\ \}$ or $\{RDS(T_j) \cap WS(T_i)\ \}$ at line 3 and 4, and checks whether $\{RS(T_j) \cap WS(T_i)\ \}$ is empty at line 4. Other lines can be computed in a constant time. Through binary searching, an element can be searched from a set A in $O(\log|A|)$. Given two sorted lists A and B, A $\cap$ B can be computed in $O(|A| + |B|)$ by scanning the two lists in the same way as the binary merge. Thus, each iteration can be computed in O(TS). As a result, the time complexity of Algorithm 3.1 is $O(mn)$ where m is DC and n is TS.

It is important to note that $O(mn)$ of concurrency control scheme implies to be a feasible algorithm for concurrency control in the following sense: (1) *MLS/OCC* scheme relies for high degree of concurrency on the hope that conflicts between transactions are rare and (2) if transaction size is small (i.e., $m > n$), algorithmic complexity is greatly dependent on the degree of concurrency.

3 Comparison with Related Works

3.1 Multiversion Orderstamp Ordering

Multiversion Orderstamp Ordering (*MVOO*) [3] is a secure scheduler which is based on multiple versions of data and a priority queue of transactions according to their access classes. *MVOO* is claimed to be secure and is able to handle write-up operations.

3.2 Orange Locking

Unlike the multiversion scheduler, the orange locking (*OOL*) scheme in [4] is for single version database, and uses locking for concurrency control. Data items can be read-locked or write-locked. When a high-level transaction wants to read lower-level data, it sets a read-down-lock on the needed data item. If a lower-level transaction wants to write a data item, it is unconditionally allowed to set a write-lock and proceed. However, if a read-down-lock is held by a higher-level transaction on this data item just locked for writing, the read-lock is changed to an orange-lock. The orange-lock indicates the potential for covert timing channel. The high-level transaction is then aborted and would have to reissue the read operation. In *OOL*, transactions wait on any conflicting lock request, and are restarted only in case of deadlocks. Each deadlock is checked for on each wait, and the transaction making the request is selected as the victim. When a transaction aborts, all locks held are released.

3.3 Comparison of Multilevel Secure Concurrency Control Schemes

We employed multiversion in *MLS/OMVTO* scheme that enjoys all the virtues of *MVOO*. Like *MVOO*, *MLS/OCC* schemes are secure and eliminate livelocks. *MLS/OCC* schemes with multiversion have two additional advantages over those of *MVOO*: (1) *MLS/OCC* schemes with multiversion do not require a totally trusted scheduler, but still need trust. (2) Moreover, the multiversion history produced by our *MLS /OCC* schemes with multiversion is equivalent to a one-serial execution in which transactions are placed in a timestamp order. As a result high-level transactions receive the most recent versions of low-level data items.

A set of *MLS/OCC* schemes allows enhanced level of concurrent executions than *OOL*, since conflicting lock modes in *OOL* can be hold simultaneously on the same data. This implies that delay suspension, which is a major problem of locking based algorithms, can be mitigated in *MLS/OCC*. However, *MLS/OCC* could lead to frequent restarts of high-level transactions due to the fact that the main strategy of conflict resolution is restart. Therefore, performance of a set of *MLS/OCC* schemes, *OOL* and *MVOO* between delay suspension and aborts should be evaluated, under a wide variety of database workloads, system configurations [5].

4 Performance Analysis

4.1 Experiment 1: Effect of Transaction Size

Fig. 2 shows the transaction throughput versus transaction size. It is interesting to observe that the *MLS/OCC*, *OOL* and *MVOO* are sensitive to the transactions with different sizes. Their performance drops as the transaction size increases.

4.2 Experiment 2: Effect of Write Ratio

There are a number of interesting observations on this experiment in Fig. 3 and 4. First, *OOL* and *MLS/OCC* still have high restart ratio. This is due to cascading aborts. Second, even though *MLS/OCC* has higher restart ratio than *OOL*, the throughput of *MLS/OCC* is better than *OOL* at low data contention.

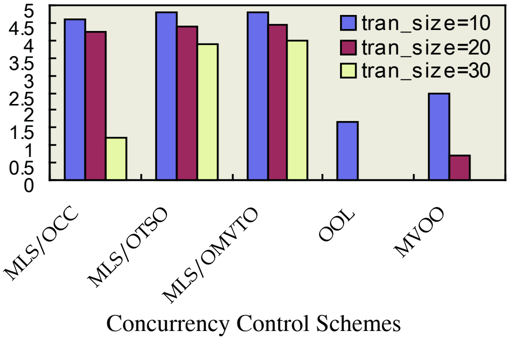
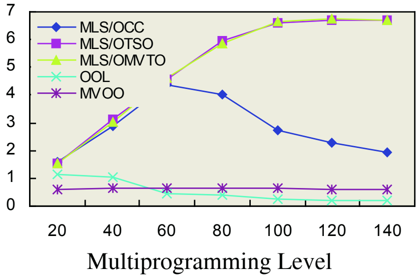


Fig. 2. Effect of Transaction Size on Throughput (*num_class*=4; *tran_size*=10, 20, 30; *MPL*=60; and *wr_op_pct*=35%)



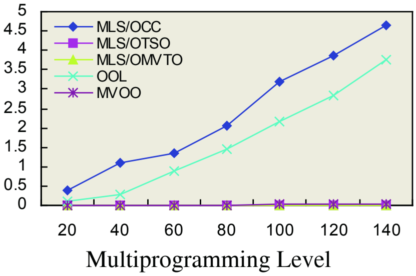


Fig. 3. Throughput: Effect of Write Ratio (*num_class*=4*, tran_size*=20, *wr_op_pct*=15%)

Fig. 4. Restart Ratio: Effect of Write Ratio (*num_class*=4*, tran_size*=20, *wr_op_pct*=15%)

5 Conclusions

In this paper, we have presented a novel covert timing channel-free concurrency control scheme that satisfies the B3 or higher level of the US TCSEC requirements in a secure database. Our simulation results substantiate our claim that a set of multilevel-secure optimistic concurrency control scheme can improve not only the significant performance but also the degree of secureness.

The *MLS/OMVTO* and *MLS/OTSO* exhibit a much better response time characteristic of most transactions than *MLS/OCC*, *OOL* and *MVOO* when write-up operation is not allowed. *MLS/OMVTO* has the best response time and throughput characteristics of all transactions. If staleness is not an issue or if average transaction sizes are small, *MLS/OMVTO* may be the secure concurrency control scheme of choice.

The *MLS/OCC* scheme causes substantially more conflicts than multiversion schemes. Hence, it is too optimistic. On the other hand, the *MVOO* and *OOL* are somewhat pessimistic. Compared to that, *MLS/OMVTO* can be applied to all degrees of secureness, namely *covert channel-freeness* and *starvation-freeness* introduced by the version idea from *MVOO*. If conflicts are very rare, *MLS/OCC* outperformed the other four algorithms by a large amount. An interesting observation is that transaction size parameter is the most sensitive factor for all concurrency control schemes.

Acknowledgements

This work was supported by a grant No.R12-2003-004-00009-0 from Ministry of Commerce, Industry and Energy.

References

1. S. Kang and S. Moon, “Read-Down Conflict-Preserving Serializability as A Correctness Criterion for Multilevel-Secure Optimistic Concurrency Control: *CSR/RD*,” Journal of System Architecture, Vol. 46, pp. 889-902, 2000.
2. H. T. Kung and J. T. Robinson, “On Optimistic Methods for Concurrency Control,” ACM Trans. Database System, Vol. 6, No. 2, pp. 213-226, June 1981.
3. T. F. Keefe and W. T. Tsai, “Multiversion Concurrency Control for Secure Database Systems,” Proc. IEEE Computer Society Symposium on Security and Privacy, pp. 369-383, May 1990.
4. J. McDermott and S. Jajodia, “Orange-Locking: Channel-Free Database Concurrency Control via Locking,” C. E. Landwehr, Database Security VI: Status and Prospects, North-Holland, pp. 262-274, 1993.
5. M. Carey and M. Stonebraker, “The Performance of Concurrency Control Algorithm for Database Management Systems,” Proc. of the 10th VLDB Conf., pp. 107-118, 1984.

Secure Password Authentication for Keystroke Dynamics

YeongGeun Choe and Soon-Ja Kim

Graduate school of Electronic Engineering, Computer Networks Lab.,
Kyungpook National University,1370, Sankyuk-dong, Buk-gu, Daegu, Korea
cygmj@korea.com

Abstract. Keystroke dynamics is an intelligent data processing technique of analyzing the user's habitual typing patterns to identify him. Keystroke dynamics combined with password authentication has been widely used as a means to enhance user authentication system. However, the user authentication system's security does not rely solely on the keystroke dynamics. To guarantee a high level of security, more secure password authentication is needed. The design and development of a secure password authentication protocol for keystroke dynamics is discussed in this paper. We propose a new efficient password authentication protocol that is secure against all types of attacks considered in the paper. We also show that our two-party protocol is extended to a three-party protocol, where each user only shares a password with a trusted server. As a result, our protocols with keystroke dynamics can provide a secure and intelligent means of authentication and access control of computer users.

1 Introduction

Keystroke dynamics, also known as typing biometrics is an intelligent data processing technique that analyzes the way a user types at a terminal by monitoring the keyboard inputs in attempt to identify users based on their habitual typing rhythm patterns. Features commonly used to describe a user's typing pattern are latencies between successive keystrokes, duration of each keystroke, finger placement, pressure applied on the keys and overall typing speed. Keystroke dynamics based on assumption that different people type in uniquely characteristic manners is combined with password authentication to enhance user authentication system. That's one reason why the keystroke dynamics functions in conjunction with a conventional password authentication. That way, the user authentication system's security does not rely solely on the keystroke dynamics. To guarantee a high level of security, more secure password authentication is needed.

In this paper, we focused on designing a secure password authentication protocol that can be used for keystroke dynamics, in all those specific applications where a high level of security is needed.

Since Bellovin and Merritt first proposed the Encrypted Key Exchange(EKE) protocol in 1992[1], many password authentication protocols have been proposed. A-EKE[2], B-SPEKE[3, 4], SRP[5], PAK[6, 7, 8] and AMP[9] are all representative verifier-based protocols that use an asymmetric model with a password for the user and a verifier for the server.

However, in scenarios where a user wants to communicate with many other users, key management becomes inconvenient when two users, where neither is the server, must mutually share verifiers. In 1995, Steiner, Tsudik and Waidner presented a

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 317–324, 2005.

three-party EKE(hereafter referred to as STW-3PEKE) protocol, which resulted in many two-party protocols being extended to three-party protocols [10, 11, 12, 13, 14, 15, 16, 17].

In this paper, we propose a new efficient password authentication protocol that is secure against all types of attacks considered in the paper. We also show that our two-party protocol is extended to a three-party protocol, where each user only shares a password with a trusted server.

The remainder of this paper is organized as follows: Section 2 proposes the new protocol, and Section 3 analyzes its security. Section 4 then extends the proposed 2-party protocol to a 3-party protocol and analyzes its security and efficiency. Some concluding remarks are given in Section 5.

2 New Protocol

This section presents a new efficient password authentication protocol. Table 1 show the notation used throughout this paper.

Table 1. Notations

Symbol	Meaning
A, B	Alice and Bob's identifier
p_A, p_B	Alice and Bob's password
v_A, v_B	Alice and Bob's verifier derived from password
p, q	large prime-numbers($p = 2q + 1$)
Z_p^*	multiplicative group
g	primitive element of Z_p^*
x, y	random variables
K	session key
h()	one-way hash function

The security of our protocols is based on the *Discrete logarithm problem* and *Diffie-Hellman problem*[18], which are infeasible to solve in polynomial time.

Let p, q be sufficiently large prime numbers such that $p = 2q+1$, and let Z_q be a subgroup of Z_p^* of order q, while g is a primitive element in Z_p^*.

Protocol setup. The user Alice selects a password *p_A*, then the sever computes the verifier $v\_A = g^{-h(p\_A)}$ mod p, which is derived from the password *p_A*, along with the user's identifier *A* using his only known secret key and stores it.

Protocol description. The protocol execution step establishes a communication channel to receive and transfer the data securely between the user Alice and the server. The proposed password authentication protocol is shown in Figure 1.

Step1. Alice chooses a random exponent x, keeps it secret, computes $G_1 = g^{x+h(p\_A)}$ mod p, then sends G_1 and her identifier *A* to the server as an initial request. After receiving Alice's request, the server looks up Alice's password entry and fetches her verifier *v_A*.

Step 2. The server also chooses a random exponent y, keeps it secret, computes $G_2 = g^y$ mod p, then sends G_2 to Alice.

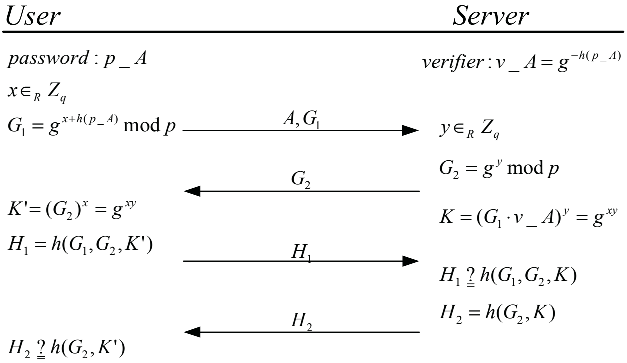

User
Server
$password : p\_A$
$verifier : v\_A = g^{-h(p\_A)}$
$x \in_R Z_q$
$G_1 = g^{x+h(p\_A)} \bmod p$
A, G_1
$y \in_R Z_q$
$G_2 = g^y \bmod p$
G_2
$K' = (G_2)^x = g^{xy}$
$K = (G_1 \cdot v\_A)^y = g^{xy}$
$H_1 = h(G_1, G_2, K')$
H_1
$H_1 \overset{?}{=} h(G_1, G_2, K)$
$H_2 = h(G_2, K)$
H_2
$H_2 \overset{?}{=} h(G_2, K')$

Fig. 1. The Proposed 2-Party Protocol

Step 3. Alice and the server compute a common exponential value $K = g^{xy} \bmod p$ using the values available to each of them. If Alice's password *p_A* entered in Step 1 matches the one she originally used to generate *v_A*, then both values of *K* will match. Alice then computes $H_1 = h(G_1, G_2, K')$ and sends it to the server as evidence that she has the correct session key.

Step 4. Next, the server computes $h(G_1, G_2, K)$ itself to verify a match with H_1 sent by Alice. The server then computes $H_2 = h(G_2, K)$ and sends it to Alice as evidence that the server also has the correct session key. Alice then verifies H_2 herself, accepting only if it matches the server's value.

3 Security and Efficiency Analysis

We show that our 2-party protocol is secure against the well-known attacks.

Let $Pr[Solve_{DLP}(k)]$ and $Pr[Solve_{DLP}(k)]$ be the probability that adversary $\mathcal{A}$ can solve the *Discrete logarithm problem* and the *Diffie-Hellman problem*. We denote the advantage of adversary, Adv_{A} as the probability that $\mathcal{A}$ correctly distinguishes the session key[19, 20, 21].

A password authentication protocol is said to be secure if for every dictionary *D* and every polynomial-time adversary $\mathcal{A}$,

$$Adv_{\mathrm{A}}(k) < 1/|D| + \varepsilon(k),$$

where $|D|$ is a dictionary size of passwords[21].

1. *Replay attack*, in which the adversary could launch it easily by replaying an eavesdropped message during a past session, is negligible because $G_1(=g^{x+h(p\_A)})$ should include an ephemeral parameter *x* of Alice, while the others such as $H_1(=h(G_1, G_2,K'))$, $G_2(=g^y)$ and $H_2(=h(G_2, K))$ should include ephemeral parameters *x*, *y* of both parties in the corresponding session. The only way to find these parameters is to obtain the ephemeral parameters *x* and *y* from the eavesdropped messages, G_1 and G_2. But, this is a *Discrete logarithm problem* and considered computational infeasible.

Therefore, an adversary can not impersonate both parties. It means that Adv_{A}^{RA} is bounded by $Pr[Solve_{DLP}(k)]$.

2. A *man-in-the-middle attack* is an active attack in which the adversary is able to read, insert, and modify at will, messages between two parties without either party knowing that the link between them has been compromised. The adversary can imitate Server when talking to Alice and imitate Alice when talking to Server. This man-in-the-middle attack is infeasible without knowing the password *p_A* or verifier *v_A*. If an adversary modifies, reflects, or replays G_1, G_2, H_1, and H_2, the attack will be detected, because the verification of H_1 and H_2 confirm both the correctness of G_1, G_2, and the session key K.

3. *Password-guessing attack*(or *dictionary attack*), where the adversary uses a dictionary of frequently selected passwords, can be divided into two types, *on-line and off-line password-guessing attack*. On-line password-guessing attack is detectable, and can be easily prevented, based on restricting the number of authentication fails. If $\mathcal{A}$ is rejected R times, the success probability of on-line password guessing attack is bounded by $1/(|D| - R)$.

The simplest way of protecting against an off-line guessing attack is to increase the computational load required to derive the password. With the proposed protocol, an off-line guessing attack is infeasible, as the adversary cannot solve G_1 for guessed passwords without knowing the high entropy parameter x. Even if an adversary intercepts the messages, G_1, G_2, H_1, and H_2, there is still no way to confirm the correctness of the guessed password *p_A'*, as no useful information about the password *p_A* or ephemeral parameter x is revealed during a session. The adversary $\mathcal{A}$ can mount a off-line guessing attack if it solves the *Discrete Logarithm problem*. So Adv_{A}^{Off} is bounded by $Pr[Solve_{DLP}(k)]$.

4. A *Denning-Sacco attack* occurs when an adversary captures the session key K from an eavesdropped session and uses it to either impersonate the user directly or conduct a brute-force search against the user's password[2, 22, 23]. Yet, even if a past session key $K = g^{xy} \bmod p$ has been revealed, the adversary must still solve the *Discrete logarithm problem* of $g^{x+h(p\_A)}$. Therefore, the proposed protocol is secure against this type of attack. Hence Adv_{A}^{DS} is bounded by $Pr[Solve_{DLP}(k)]$.

Also, our 2-party protocol satisfies the property of *perfect forward secrecy*. Even when the password *p_A* itself is compromised, an adversary can not compute past session keys without solving the computation *Diffie-Hellman problem* of session key $K = g^{xy}$ from given $G_1(= g^{x+h(p_A)})$and $G_2(= g^y)$. Therefore Adv_{A}^{PFS} is bounded by $Pr[Solve_{DHP}(k)]$.

From the analysis above, we conclude that

$$Adv_{\mathrm{A}}(k) < 1/(|D| - R) + Adv_{\mathrm{A}}^{RA} + Adv_{\mathrm{A}}^{Off} + Adv_{\mathrm{A}}^{DS} + Adv_{\mathrm{A}}^{PFS} + \varepsilon(k).$$

Since Adv_{A}^{RA}, Adv_{A}^{Off}, Adv_{A}^{DS} and Adv_{A}^{PFS} have been shown to be negligible, the proposed protocol is secure.

In Table 2, the efficiency of the proposed protocol was compared to that of other related protocols, such as A-EKE, B-SPEKE, SRP, PAK-X/R/RY, AMP as regards several factors, including the number of encryptions, hash functions, exponentiations and random numbers, and protocols steps.

As shown in Table 2, the proposed 2-party protocol is more efficient than the existing protocols.

Table 2. Comparison of Efficiency of 2-Party Protocol

	Pass	Encryption		Hash func.		Exponent		Random No.	
		User	Server	User	Server	User	Server	User	Server
A-EKE	5	3	3	1	1	4	4	1	1
B-SPEKE	4	×	×	1	1	3	4	1	2
SRP	4	×	×	3	2	3	3	1	1
PAK-X	3	×	×	4	4	4	4	1	2
PAK-R	3	×	×	3	3	3	3	2	1
PAK-RY	3	×	×	3	3	4	5	3	1
AMP	4	×	×	3	3	2	3	1	1
Proposed	**4**	×	×	**3**	**2**	**2**	**2**	**1**	**1**

4 Extended 3-Party Protocol

In this section, we propose a three-party protocol based on our two-party password authentication protocol and compare the related protocols.

4.1 3-Party Protocol

The two-party protocol proposed in Section 2 can be extended to a three-party protocol, as shown in Figure 2, which is achieved by letting the trusted third-party(server) act as a relay.

Step 1. Alice computes X_A using her password p_A, chooses a random exponent x, keeps it secret, and computes $G_A = g^x$ mod p and $G_1 = G_A\square X_A = g^{x+h(p\_A)}$ mod p. Alice then sends the ID's A, B, and G_1 as an initial request to the server. After receiving Alice's request, the server looks up the password database and fetches Alice and Bob's verifiers, v_A and v_B, respectively.

Step 2. The server computes $G_2 = G_1\square v\_A\square v\_B = g^{x-h(p\_B)}$ mod p and sends G_2 and A to Bob.

Step 3. Bob computes X_B using his password p_B, chooses a random exponent y, keeps it secret, and computes $G_B = g^y$ mod p and $G_3 = G_B\square X_B = g^{x+h(p\_B)}$ mod p. These messages, G_B and G_3, can be computed previously, regardless of receiving message 2. After receiving the server's request, Bob then sends the ID's A, B, and G_3 to the server. While waiting for message 5, Bob computes $T_B = G_2\square X_B$ and $K = (T_B)^y = g^{xy}$ mod p.

Step 4. The server computes $G_4 = G_3\square v\_A\square v\_B = g^{x-h(p\_A)}$ mod p and sends G_4 and B to Alice.

Step 5. Alice computes $T_A = G_4\square X_A$, $K\square = (T_A)^x = g^{xy}$ mod p and $H_1 = h(G_A, T_A, K\square)$, then sends H_1 to Bob.

Step 6. Bob computes $h(T_B, G_B, K)$ himself and verifies that it matches H_1 sent to him by Alice. Bob then computes $H_2 = h(G_B, K)$ and sends it to Alice as evidence that he also has the correct session key. Alice also verifies H_2 herself, accepting only if it matches Bob's value.

4.2 Security and Efficiency Analysis

Just like its 2-party counterpart, the 3-party protocol is also resistant to the various attacks considered in Section 3. Replay attack is negligible, as the transmitted mes-

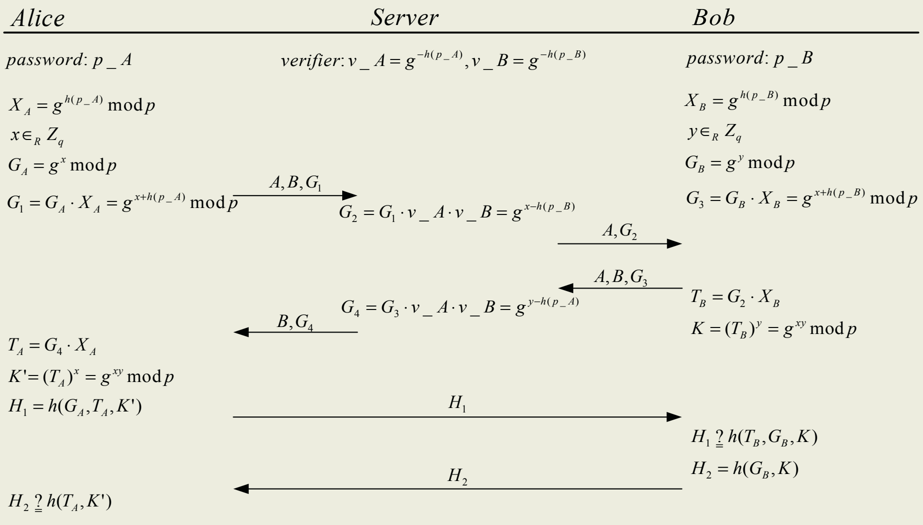

Alice
Server
Bob
password: p_A
verifier: $v\_A = g^{-h(p\_A)}, v\_B = g^{-h(p\_B)}$
password: p_B
$X_A = g^{h(p\_A)} \bmod p$
$x \in_R Z_q$
$G_A = g^x \bmod p$
$G_1 = G_A \cdot X_A = g^{x+h(p\_A)} \bmod p$
A, B, G_1
$G_2 = G_1 \cdot v\_A \cdot v\_B = g^{x-h(p\_B)}$
A, G_2
$X_B = g^{h(p\_B)} \bmod p$
$y \in_R Z_q$
$G_B = g^y \bmod p$
$G_3 = G_B \cdot X_B = g^{x+h(p\_B)} \bmod p$
A, B, G_3
$T_B = G_2 \cdot X_B$
$K = (T_B)^y = g^{xy} \bmod p$
$G_4 = G_3 \cdot v\_A \cdot v\_B = g^{y-h(p\_A)}$
B, G_4
$T_A = G_4 \cdot X_A$
$K' = (T_A)^x = g^{xy} \bmod p$
$H_1 = h(G_A, T_A, K')$
H_1
$H_1 \overset{?}{=} h(T_B, G_B, K)$
$H_2 = h(G_B, K)$
H_2
$H_2 \overset{?}{=} h(T_A, K')$

Fig. 2. The Extended 3-Party Protocol

sages, G_1, G_2, G_3, G_4, H_1, and H_2, must include the ephemeral parameters for the two parties, and obtaining these ephemeral parameters corresponds to solving a *Discrete logarithm problem*. Meanwhile, a man-in-the-middle attack is infeasible without knowing the user's passwords *p_A* and *p_B* or verifiers *v_A* and *v_B*. An off-line guessing attack will also not work on the proposed protocol, because the adversary cannot solve G_1, G_2, G_3, and G_4 for guessed passwords. Even if a past session key is compromised, the adversary must still solve the *Discrete logarithm problem* of $g^{x+h(p\_A)}$ or $g^{y+h(p\_B)}$ to impersonate the user, making the proposed protocol secure against a Denning-Sacco attack. The proposed protocol also satisfies the property of perfect forward secrecy, as a compromised password does not allow an adversary to determine the session key *K* for past sessions.

The efficiency of the proposed protocol was compared to that of other related protocols, such LSH-3PEKE, LSSH-3PEKE[14, 16], as regards several factors, including the number of exponentiations, public-key en/decryption, symmetric en/decryption, PRF operation, hash, random numbers, and protocols steps, as shown in Table 3.

Table 3. Efficiency Comparison

	LSH-3PEKE			LSSH-3PEKE			**Proposed protocol**		
	A	B	S	A	B	S	**A**	**B**	**S**
Exponentiation	2(7)	2(7)	0(6)	3	3	4	**3**	**3**	**0**
public-key en/decryption	1/0	1/0	0/2	0	0	0	**0**	**0**	**0**
symmetric en/decryption	2	2	2	1	1	2	**0**	**0**	**0**
PRF operation	0	0	0	3	3	4	**0**	**0**	**0**
hash	1	1	2	2	2	2	**2**	**2**	**0**
random numbers	2	3	0	1	1	2	**1**	**1**	**0**
protocol steps	5			7			**6**		

It should be noted that the best-known nonmalleable encryption scheme[24] needs 5 and 3 exponentiations per encryption and decryption, respectively, while LSH-3PEKE and LSSH-3PEKE are password-equivalent protocols.

Therefore, the proposed 3-party protocol clearly exhibited a lower computational cost than the existing protocols.

5 Conclusion

Although keystroke dynamics is combined with password authentication to enhance user authentication, more secure password authentication is needed to guarantee a high level of security.

In this paper, we focused on designing a secure password authentication protocol that can be used for keystroke dynamics. As a result, when using the proposed protocol, an adversary cannot feasibly learn either the user's password or session key by monitoring communications and compromising passwords and session keys. Also, the proposed 2-party protocol can be extended to a 3-party protocol, where each user shares a password only with a trusted server without sharing a password between users.

Clearly, our proposed password authentication protocols combined with keystroke dynamics can provide a secure and intelligent means of authentication and access control of computer users in terms of ease of use, improved security and performance.

References

1. S. Bellovin and M. Merritt. Encrypted key exchange: password-based protocols secure against dictionary attacks. In IEEE Symposium on Research in Security and Privacy, pp. 77-84, 1992
2. S. Bellovin and M. Merritt. Augmented encrypted key exchange: a password-based protocols secure against dictionary attacks and password-file compromise. In ACM Conference on Computer and Communications Security, pp. 244-250, 1993
3. D. Jablon. Strong password-only authenticated key exchange. ACM Computer Communications Review, Vol. 26, no. 5, pp. 5-26, 1996
4. D. Jablon. Extended password key exchange protocols immune to dictionary attacks. In WETICE'97 Workshop on Enterprise Security, pp. 248-255, 1997
5. T. Wu. Secure remote password protocol. In Network and Distributed System Security Symposium Conference Proceedings, 1998
6. V. Boyko, P. MacKenzie and S. Patel. Provably secure password authenticated key exchange using Diffie-Hellman. In Eurocrypt '00, pp.156-171, 2000
7. P. MacKenzie. More Efficient Password-Authenticated Key Exchange. In CT-RSA 2001, pp. 361-377, 2001
8. P. MacKenzie. The PAK suites: Protocols for Password-Authenticated Key Exchange. 2002, available from http://grouper.ieee.org/groups/1363/passwdPK/contributions.html#Mac02
9. T. Kwon. Authentication and Key agreement via Memorable Passwords. In Network and Distributed System Security Symposium Conference Proceedings, 2001
10. T. Kwon, M. Kang and J. Song. An Adaptable and Reliable Authentication Protocol for Communication Networks. In Proceedings of IEEE INFOCOM'97, pp. 737-744, 1997
11. T. Kwon, M. Kang, S. Jung and J. Song. An Improvement of the Password-based Authentication protocol(K1P) on Security against Replay Attacks. In IEICE Transactions on Communications, E82-B(7), pp. 991-997, 1999

12. Y. Ding and P. Horster. Undetectable On-line Password Guessing Attacks. ACM Operating Systems Review, Vol. 29, no. 4, pp. 77-86, 1995
13. C.-L. Lin, H.-M. Sun, and T. Hwang. Three party encrypted key exchange: Attacks and a solution. ACM Operating Systems Review, Vol. 34, no. 4, pp. 12-20, 2000
14. C.-L. Lin, H.-M. Sun, M. Steiner and T. Hwang. Three-party encrypted key exchange Without Server Public-Keys. In IEEE, Communications Letters, Vol. 5, no. 12, pp. 497-499, 2001
15. L. Gong, M. Lomos, R. Needham. Protecting Poorly Chosen Secrets from Guessing Attacks. In IEEE Journal on Selected Areas in Communications, Vol. 11, no. 5, pp. 648-656, 1993
16. M. Steiner, G. Tsudik, M. Waidner. Refinement and Extension of Encrypted Key Exchange. In ACM Operating Systems Review, Vol. 29, no. 3, pp. 22-30, 1995
17. L. Gong. Optimal Authentication Protocols Resistant to Password Guessing Attacks. In 8th IEEE Computer Security Foundations Workshop, pp. 24-29, 1995
18. W. Diffie and M. Hellman. New directions in cryptograpy. In IEEE Transactions on Information Theory, Vol. 22, no. 6, pp. 644-654, 1976
19. Shai Halevi and Hugo Krawczyk. Public-key cryptography and password protocols. In ACM Transactions on Information and System Security, Vol. 2, no. 3, pp. 230-268, 1999
20. M. Bellare, D. Pointcheval, and P. Rogaway. Authenticated Key Exchange Secure Against Dictionary Attacks. In Eurocrypt 2000, pp. 139-155, 2000
21. R. Gennaro and Y. Lindell. A Framework for Password-Based Authenticated Key Exchange. In Eurocrypt 2003, pp. 524-543, 2003
22. D. Denning and G. Sacco. Timestamps in key distribution protocols. Communications of the ACM, Vol. 24, no. 8, pp. 533-536, 1981
23. Y. Yacobi. A key distribution paradox. In Crypto ’90, pp. 268-273, 1990
24. R. Cramer and V. Shoup. A practical public key cryptosystem provably secure against adaptive chosen ciphertext attack. In Crypto ’98, pp. 13-25, 1998

The Efficient Multipurpose Convertible Undeniable Signature Scheme

Sung-Hyun Yun[1] and Hyung-Woo Lee[2]

[1] Division of Information and Communication Engineering, Cheonan University,
Anseo-dong, Cheonan, 330-704, Korea
shyoon@infocom.cheonan.ac.kr

[2] Dept. of Software, Hanshin University, Osan, Gyunggi, 447-791, Korea
hwlee@hs.ac.kr

Abstract. The digital signature can be verified and disavowed only with cooperation of the signer in undeniable signature scheme. A signed confidential document of a company can be copied and delivered to a rival company. If a conventional signature scheme is used to sign the document, it can be confirmed as authentic by verifying the signature without the signer's cooperation. However, if the company doesn't want the document to be verified as authentic by the rival company, it is recommended to use the undeniable signature scheme. Convertible undeniable signature scheme has additional property that the signer can convert undeniable signature to the ordinary one. The document signed by undeniable signature scheme that is no longer confidential can be opened to public use by converting the signature to the ordinary one. In this study, the efficient multipurpose convertible undeniable signature scheme based on El-Gamal signature scheme is proposed. The proposed scheme satisfies undeniable property and can convert undeniable signature to the ordinary one. The number of public keys and signatures are less than those of Boyar's convertible signature scheme[4]. It also reduces the number of communication steps of the signature confirmation protocol.

1 Introduction

Undeniable signature scheme is a method that signature can not be verified without signer's cooperation. Convertible undeniable signature scheme has additional property that undeniable signature can be converted to the ordinary one by releasing partial secret information. It has many applications where an ordinary digital signature scheme can not be applied to.

In on-line sales of digital contents, an owner of the digital contents wants to know whether a distributor sales the contents to customers fairly. In this case, the owner can satisfy on-line sales model if the model provides the mechanism which the customer can not buy the digital contents without help of the owner. The digital copyright generated by undeniable signature scheme differs from the ordinary copyright in that the customer can not distinguish between valid and invalid copyright without help of the copyright owner. Only the original owner can confirm this copyright as authentic to the customer.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 325–331, 2005.

However, in customers' position, they want to verify signature on the digital contents by themselves after purchasing the contents. In this case, customers can satisfy if the copyright owner makes the undeniable signature on the contents to the ordinary one. Convertible undeniable signature scheme can be used to solve above mentioned problems.

In this study, we show how to use undeniable concept of Chaum[1] and El-Gamal signature scheme[3] to obtain new convertible undeniable signature scheme. Modified El-Gamal signature scheme based on discrete logarithms is proposed. The proposed scheme satisfies undeniable property[1] and can convert the undeniable signature to the ordinary one. The size of the public key and the signature is less than that of Boyar's convertible scheme[4]. It also reduces the number of communication steps between signer and verifier.

Boyar's convertible undeniable signature scheme is described in section 2. The proposed efficient convertible undeniable signature scheme is presented in section 3 and analyzed in section 4. Section 5 compares the proposed scheme with the Boyar's scheme. conclusion and applications of our works are presented in section 6.

2 Related Works

In this section we review existing Boyar's convertible undeniable signature scheme [4]. The security of Boyar's scheme is based on the difficulty of solving discrete logarithms over $GF(p)$. The signer selects secret keys $x, z \in Z_q$, message $m \in Z_q$ and generates public keys $y \equiv g^x \ (mod\ p),\ u \equiv g^z \ (mod\ p)$.

(1) Signature Generation Prorocol

Step 1. A signer chooses random number t and makes the message $M \equiv g^t tzm$ $(mod\ q)$.

Step 2. The signer generates El-Gamal signautre (r, s) on the M.

$$g^M \equiv y^r r^s \ (mod\ p)$$

Step 3. The signer sends (g^t, r, s, m) to the verifier.

(2) Signature Confirmation Protocol

The signer and the verifier generate $(w \equiv g^{t \cdot g^t \cdot m} \ (mod\ p),\ v \equiv y^r r^s \ (mod\ p))$.

Step 1. The verifier chooses random numbers (a, b) and generates the challenge $ch \equiv w^a g^b \ (mod\ p)$. The verifier sends ch to the signer.

Step 2. The signer chooses random number r and generates responses $(h_1 \equiv ch^r \ (mod\ p),\ h_2 \equiv {h_1}^z \ (mod\ p))$. The signer sends (h_1, h_2) to the verifier.

Step 3. The verifier sends (a, b) to the signer.

Step 4. The signer authenticates the challenge ch and sends r to the verifier.

Step 5. The verifier authenticates (h_1, h_2) as follows.

$$h_1 \equiv (w^a g^b)^r \ (mod\ p), \quad h_2 \equiv (v^a u^b)^r \ (mod\ p)$$

(3) Signature Conversion Protocol
The signer can transform the undeniable signatures to the ordinary ones by releasing z. If signer wants to convert particular undeniable signature, corresponding t is released. When t is released, signature verification is proceeded as follows.

$$(u^{m\cdot g^t})^t \equiv y^r r^s \ (mod\ p)$$

3 The Proposed Scheme

In this section, undeniable signature scheme based on El-Gamal scheme is proposed. The proposed scheme satisfies undeniability and can convert undeniable signature to the ordinary one.

We modify the El-Gamal signature equation as follows.

$$k(m+s) \equiv xr \ (mod\ p-1) \quad (3.1)$$

A signer's private key x, public key y and message m are defined as follows.

$$y \equiv g^x \ (mod\ p), \quad x, m \in Z_{p-1}$$

3.1 Signature Generation Protocol

Step 1. A signer chooses random number k and keeps it secret. k and $p-1$ are relatively prime numbers. One way hash function h is used to hash the message m. Let the hash result $m_h = h(m, hpr)$ be a primitive root of mod p by adjusting hash parameter hpr.
Step 2. The signer generates public parameter $r \equiv {m_h}^k \ (mod\ p)$.
Step 3. The signer chooses s satisfying following equation 3.2.

$$ks \equiv xr - km_h \ (mod\ p-1) \quad (3.2)$$

Step 4. The signer sends (m_h, hpr, r, s) to the verifier.

3.2 Signature Confirmation Protocol

Step 1. The verifier chooses random numbers (a, b) and generates the challenge $ch \equiv r^{a(m_h+s)} \cdot y^{rb} \ (mod\ p)$.
Step 2. The verifier sends ch to the signer.
Step 3. The signer generates the response $rsp \equiv ch^{x^{-1}} \ (mod\ p)$ and sends it to the verifier.
Step 4. The verifier authenticates the signature (r, s) as follows.

$$rsp \equiv m_h^{ra} \cdot g^{rb} \ (mod\ p) \quad (3.3)$$

If the signature is invalid or the signer denies valid signature, equation 3.3 can not be hold. Undeniability of the proposed scheme is proved in section 4. In order

to convert undeniable signature to the ordinary one, the following additional steps are required.
Step 5. The verifier sends (a, b) to the signer.
Step 6. The signer authenticates the challenge $ch \equiv r^{a(m_h+s)} \cdot y^{rb}\ (mod\ p)$.
Step 7. The signer converts undeniable signature to the ordinary one by releasing g^k.

$$g^{k(m_h+s)} \equiv y^r\ (mod\ p)$$

3.3 Disavowal Protocol

If the verifier fails to authenticate undeniable signature in step 4 of the signature confirmation protocol, he/she can perform following disavowal protocol to identify whether the signature is invalid or the signer denies valid signature. The security of the proposed disavowal protocol is analyzed in theorem 2 and 3.
Step 5: The verifier chooses random numbers (c, d) and generates the challenge $ch' \equiv r^{c(m_h+s)} \cdot y^{rd}\ (mod\ p),\ ad \neq bc.$
Step 6: The signer sends the response $rsp' \equiv ch'^{x^{-1}}\ (mod\ p)$ to the verifier.
Step 7: The verifier generates following discrimination equations.

$$R_1 \equiv (rsp \cdot g^{-rb})^c\ (mod\ p), \quad R_2 \equiv (rsp' \cdot g^{-rd})^a\ (mod\ p)$$

Step 8: The verifier compares R_1 with R_2 as follows.
$R_1 = R_2$: invalid signature, $R_1 \neq R_2$: the signer denies valid signature

4 Security Analysis

We analyze the undeniability of the proposed scheme. Theorem 1 shows that signer can not generate valid response for an invalid signature. Theorem 2 and 3 are proofs of the proposed disavowal protocol.

Definition 1 *The valid signature* (r, s) *and the invalid signature* (r', s) *are defined as follows.*

- *valid signature* (r, s)

$$k(m_h + s) \equiv xr\ (mod\ p-1), \quad r \equiv {m_h}^k\ (mod\ p)$$

- *invalid signature* (r', s)

$$k'(m_h + s) \equiv x'r'\ (mod\ p-1), \quad r' \equiv {m_h}^{k'}\ (mod\ p), \quad x \neq x'$$

Theorem 1 *A signer can not generate valid response for an invalid signature.*

Proof: The followings are challenges and responses for an invalid signature (r', s). We assume that (a, b) and (a', b') generate same challenge.

- Challenge ch and response rsp with (a, b).

$$ch \equiv {m_h}^{k'(m_h+s)a} \cdot g^{xr'b}\ (mod\ p),\ rsp \equiv {m_h}^{r'a} \cdot g^{r'b}\ (mod\ p)$$

- Challenge ch' and response rsp' with (a', b').

$$ch' \equiv {m_h}^{k'(m_h+s)a'} \cdot g^{xr'b'} \ (mod\ p),\ rsp' \equiv {m_h}^{r'a'} \cdot g^{r'b'} \ (mod\ p)$$

The following equations are obtained since two pairs (a, b) and (a', b') yield same challenge.

$${m_h}^{k'(m_h+s)(a-a')} \equiv g^{xr'(b'-b)} \ (mod\ p) \quad (4.1)$$

$${m_h}^{r'(a-a')} \equiv g^{r'(b'-b)} \ (mod\ p) \quad (4.2)$$

Equation 4.3 is obtained by replacing the part of right side of equation 4.1 with left side of equation 4.2.

$${m_h}^{k'(m_h+s)(a-a')} \equiv {m_h}^{xr'(a-a')} \ (mod\ p) \quad (4.3)$$

In definition 1, (r', s) is an invalid signature and holds $k'(m_h + s) \neq xr'$ $(mod\ p-1)$. Equation 4.3 can not be hold in this case. Therefore signer can not generate valid response for an invalid signature more than once.

Theorem 2 *The proposed disavowal protocol can prove that whether signer answers improperly or not for a valid signature.*

Proof: The valid signature (r, s) is defined in definition 1. The verifier generates the challenge $ch \equiv r^{a(m_h+s)} \cdot y^{rb} \ (mod\ p)$.

The signer generates the response r_1 improperly in order to deny valid signature as follows.

$$r_1 \equiv ch^t \equiv {m_h}^{xrat} \cdot g^{xrbt} \not\equiv {m_h}^{ra} \cdot g^{rb} \ (mod\ p)$$

Since the response r_1 is not congruent to ${m_h}^{ra} \cdot g^{rb}$ $(mod\ p)$, the verifier performs following disavowal protocol with new challenge values (c, d).

$$ch' \equiv r^{c(m_h+s)} \cdot y^{rd} \ (mod\ p),\ r_2 \equiv {ch'}^t \equiv {m_h}^{xrct} \cdot g^{xrdt} \not\equiv {m_h}^{rc} \cdot g^{rd} \ (mod\ p)$$

The verifier generates following discrimination equations with r_1 and r_2.

$$R_1 \equiv (r_1 \cdot g^{-rb})^c \equiv {m_h}^{xract} \cdot g^{(xrbt-rb)c} \ (mod\ p)$$

$$R_2 \equiv (r_2 \cdot g^{-rd})^a \equiv {m_h}^{xract} \cdot g^{(xrdt-rd)a} \ (mod\ p)$$

In the proposed disavowal protocol, if the signer answers improperly for the valid signature then R_1 is not equal to R_2. The following equations show that the signer answers improper responses.

$$(xrbt - rb)c \not\equiv (xrdt - rd)a \ (mod\ p-1)$$

Theorem 3 *The proposed disavowal protocol can prove that whether signature is valid or not.*

Proof: Invalid signature (r', s) is defined in definition 1. The verifier generates the challenge $ch \equiv {r'}^{a(m_h+s)} \cdot y^{r'b} \ (mod\ p)$.

The signer generates the response r_1 for an invalid signature as follows.

$$r_1 \equiv ch^{x^{-1}} \equiv {m_h}^{k'a(m_h+s)x^{-1}} \cdot g^{r'b} \neq {m_h}^{r'a} \cdot g^{r'b}\ (mod\ p)$$

Since the response r_1 is not congruent to ${m_h}^{r'a} \cdot g^{r'b}$ $(mod\ p)$, the verifier performs following disavowal protocol.

$$ch' \equiv {r'}^{c(m_h+s)} \cdot y^{r'd}\ (mod\ p)$$

$$r_2 \equiv {ch'}^{x^{-1}} \equiv {m_h}^{k'c(m_h+s)x^{-1}} \cdot g^{r'd}\ (mod\ p)$$

The verifier generates following discrimination equations with r_1 and r_2. In the proposed disavowal protocol, if the signer answers properly for an invalid signature then R_1 is equal to R_2. The following equations show that signature is invalid.

$$R_1 \equiv (r_1 \cdot g^{-r'b})^c \equiv {m_h}^{k'acx^{-1}(m_h+s)}\ (mod\ p)$$

$$R_2 \equiv (r_2 \cdot g^{-r'd})^a \equiv {m_h}^{k'acx^{-1}(m_h+s)}\ (mod\ p)$$

5 Discussion

Table 1 shows the comparative results between Boyar's scheme and the proposed scheme. Considering the number of communication steps between the signer and the verifier in the signature confirmation protocol, Boyar's scheme based on simultaneous discrete logarithms[4] requires 4 steps. Proposed scheme requires only 2 steps as like that of Chaum's scheme[1]. Signature size of [1] is less than that of [4] and the proposed scheme. In order to convert undeniable signature to the ordinary one, [4] requires one more parameter than that of our scheme.

Table 1. The Comparative Results between Boyar's Scheme and The Proposed Scheme

	Boyar's Scheme	The Proposed Scheme
Communication Steps	4	2
Signature Size(bit)	3n	2n
Public Key Size(bit)	5n	3n

Boyar's scheme[4] has drawbacks in communication steps, signature size and public key size compared with those of Chaum's scheme[1]. But, [4] has property that it can convert undeniable signature to the ordinary one and can make several different signatures on the same message.

By modifying El-Gamal signature equation, the proposed convertible undeniable signature scheme can achieve good performance on communication steps and key size compared to Boyar's scheme[4].

Chaum's scheme[1] can not convert undeniable signature to the ordinary one. It always produces same signature on the same message.

6 Conclusion

In this study, convertible undeniable signature scheme based on modified ElGamal signature scheme is proposed. It can convert undeniable signature to the ordinary one and can make different signatures on the same message. The communication steps of the proposed signature confirmation protocol is less than that of Boyar's scheme[4]. Public key and signature size are also less than those of [4].

The proposed scheme satisfies undeniability and can easily be extended to the convertible undeniable multi-signature scheme. The multi-signature scheme can be used to computerize application where it requires many signers and designated verifier for the security of the application. It is best suited to electronic voting and copyright protection schemes. It can minimize the role of the trusted voting center and can provide voter-oriented election. It also can be applied to joint-copyright protection where many authors can share copyright on the digital contents.

Acknowledgments

This research was supported by Cheonan University. The second author of this paper was supported by ITRC project.

References

1. D. Chaum, "Undeniable Signatures," Advances in Cryptology, Proceedings of CRYPTO'89, Springer-Verlag, pp.212-216, 1990.
2. W. Diffie and M. Hellman. New directions in cryptography. IEEE Transactions on Information Theory, IT-22(6):472. 492, November 1976.
3. T. ElGamal. A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Transactions on Information Theory, IT-30(4):469.472, July 1985.
4. I. B. D. J. Boyar, D. Chaum and T. P. Pedersen. Convertible undeniable signatures. Advances in Cryptology, Proceedings of Crypto'90, pages 189.205, 1991.
5. M. M. Patrick Horster and H. Petersen. Blind multisignature schemes and their relevance for electronic voting. Proceedings of COMPSAC'95, pages 149.155, 1995.
6. T. P. Pedersen. Distributed provers with applications to undeniable signatures. Advances in Cryptology, Proceedings of Eurocrypt'91, LNCS 547, pages 221.242, 1991.
7. S. H. Yun, S. J. Lee, "An electronic voting scheme based on undeniable signature scheme," Proceedings of IEEE 37th carnahan conference on Security Technology, pp.163-167, 2003.

A New Efficient Fingerprint-Based Remote User Authentication Scheme for Multimedia Systems

Eun-Jun Yoon and Kee-Young Yoo*

Department of Computer Engineering, Kyungpook National University,
Daegu 702-701, Republic of Korea
ejyoon@infosec.knu.ac.kr, yook@knu.ac.kr

Abstract. In applications of multimedia communication, the remote user authentication is an important issue to protect the multimedia resources. Recently, Lin-Lai proposed an improvement on the Lee-Ryu-Yoo fingerprint-based remote user authentication scheme using smart cards to not only keep the original advantages but also enhance their security. Lin-Lai claimed that their scheme can withstand masquerade attacks and secure access control in multimedia communication. However, we have found that the scheme is still susceptible to impersonation attacks. Accordingly, the current paper demonstrates the vulnerability of the Lin-Lai scheme to impersonation attacks and presents an improved scheme with better security strength and efficiency. The computational costs of the proposed scheme are less than those of the Lin-Lai scheme.

Keyword: Security techniques, Multimedia, User authentication, Fingerprint verification, Biometrics, Smart card

1 Introduction

User authentication is an important part of security, along with confidentiality and integrity, for systems that allow remote access over untrustworthy networks like the Internet. As such, a remote password authentication scheme authenticates the legitimacy of users over an insecure channel, where the password is often regarded as a secret shared between the remote system and the user. Based on knowledge of the password, a user can use it to create and send a valid login message to a remote system to gain the right to access. Meanwhile, the remote system also uses the shared password to check the validity of the login message and authenticate the user.

In 1981, Lamport [1] proposed a remote password authentication scheme that uses a password table to achieve user authentication. However, one of the weaknesses of Lamport's scheme is that a verification table should be stored in the remote system in order to verify the legitimacy of a user. If an attacker can somehow break into the server, the contents of the verification table can be easily modified. Thus, recently, many password authentication schemes have proposed

* Corresponding author: Kee-Young Yoo
Tel.: +82-53-950-5553; Fax: +82-53-957-4846

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 332–338, 2005.

solutions using smart cards in which the verification table is no longer required in the remote system to improve security, functionality, and efficiency.

In 2002, Lee-Ryu-Yoo [2] proposed a fingerprint-based remote user authentication scheme using smart cards. Their scheme was based on the ElGamal's public key cryptosystem [3] with two secret keys. In addition, their scheme strengthened the system security by verifying the smart card owner's fingerprint. Their fingerprint verification method is based on minutiae extraction and matching [4–6]. That is, a different map of minutiae will be made when the input device takes a smart card owner's fingerprint. Then the scheme can generate a one-time random number for the ElGamal's public key cryptosystem by using the map of minutiae.

In 2004, Lin-Lai [7], however, pointed out that the Lee-Ryu-Yoo scheme can not withstand masquerade attack by using two secret keys and fingerprint verification. Furthermore, they proposed an improved scheme to enhance the security. Based on ElGamal's cryptosystem and fingerprint verification, the Lin-Lai scheme needs only to maintain one secret key, without verification tables such as a password and identity tables. They claimed that their scheme provides effective authentication and also eliminates the drawback of the Lee-Ryu-Yoo scheme.

The Lin-Lai scheme, however, is still vulnerable to impersonation attacks. Accordingly, the current study demonstrates that the Lin-Lai scheme is still susceptible to impersonation attacks, in which an attacker can easily impersonate other legal users to access the resources at a remote system, and then presents an enhancement to the scheme that offers better security strength and efficiency.

The rest of this paper is organized as follows: In Section 2, we briefly review the Lin-Lai scheme. In Section 3, we point out the some security flaws. In Section 4, we improve the Lin-Lai scheme. In Sections 5 and 6, we analyze the security and efficiency of our proposed scheme, respectively. Finally, conclusions are provided in Section 7.

2 Review of the Lin-Lai Scheme

This section briefly reviews the Lin-Lai scheme. There are three phases in the scheme including a registration phase, a login phase, and an authentication phase.

Registration Phase: Let P be a large prime number and $h(\cdot)$ a one-way hash function. U_i denotes a legal user. ID_i and PW_i denotes U_i's identity and password, respectively. Before accessing a remote system, a new user U_i must imprint his/her fingerprint minutiae and choose his/her identity ID_i and password PW_i; then he/she offers the ID_i and PW_i to the registration center. The remote system computes $ID_i' = h(ID_i \oplus PW_i)$ and $PW_i' = (ID_i')^{SK} \bmod P$, where SK is a secret key maintained by the remote system and $\oplus$ is an exclusive or operation. Then, the registration center stores $(h(\cdot), P, PW_i')$ on the U_i's smart card and delivers it to the user U_i through a secure channel. The smart card keeps the PW_i and user's fingerprint minutiae template secretly. The fingerprint information possessed by each user is be different. U_i has his/her own smart card that can authenticate his/her ownership by matching the fingerprint from the extracted minutiae.

Login Phase: Whenever a user U_i wants to login, U_i has to insert his/her own smart card into the card reader and imprint the fingerprint. Then he/she types in identity ID_i and password PW_i. If U_i passes the fingerprint verification, U_i's smart card will perform the following operation:

1. Generate a random number r using the minutiae extracted from the imprint fingerprint.
2. Compute $ID'_i = h(ID_i \oplus PW_i)$.
3. Compute $C_1 = (ID'_i)^r \bmod P$.
4. Compute $t = h(T \oplus PW'_i) \bmod (p-1)$, where T is the current timestamp of the input device.
5. Compute $M = (ID_i)^t \bmod P$.
6. Compute $C_2 = M(PW'_i)^r \bmod P$.
7. Send a message $C = (ID_i, C_1, C_2, T, PW'_i)$ to the remote system.

Authentication Phase: After a transmission delay, the system receives the message C at T', where T' is the receiving timestamp of the system. The system then performs the following operations:

1. The system checks whether the format of ID_i is correct or not. If the format is incorrect, the system rejects the login request.
2. If $|T - T'| \geq \mathit{\Delta T}$, where $\mathit{\Delta T}$ denotes the expected valid time interval for transmission delay, the system rejects the login request.
3. Verifying whether $C_2(C_1^{SK})^{-1} \stackrel{?}{=} ID_i^{h(T \oplus PW'_i)} \bmod P$, if being successful, the system accepts the login request. Otherwise, the system rejects the login request.

3 Impersonation Attack on the Lin-Lai Scheme

This section demonstrates that Lin-Lai scheme is vulnerable to impersonation attack, in which an attacker can easily impersonate other legal users to access the resources at a remote system. Suppose that an attacker has eavesdropped a valid message $C = (ID_i, C_1, C_2, T, PW'_i)$ from an open network. In the login phase, the attacker can perform the impersonation attack as follows:

1. Get $(PW'_i)^r$ by computing $C_2(ID_i^{h(T \oplus PW'_i)})^{-1} \bmod P$.
2. Compute $t^* = h(T^* \oplus PW'_i) \bmod (p-1)$, where T^* is the attacker's the current date and time for succeeding with Step 2 of the authentication phase.
3. Compute $M^* = (ID_i)^{t^*} \bmod P$.
4. Compute $C_2^* = M^*(PW'_i)^r \bmod P$.
5. Send a message $C^* = (ID_i, C_1, C_2^*, T^*, PW'_i)$ to the remote system.

When the remote system receives the message C^*, it will go into the authentication phase and performs the following checks.

1. It checks the format of the ID_i. Of course, it is correct.
2. Then, it checks the time is valid or not. Because $|T^* - T'| \geq \Delta T$, where T' is the received timestamp of message C^*, the system will accept this check.
3. Finally, it compares

$$C_2^*(C_1^{SK})^{-1} \overset{?}{=} ID_i^{h(T^* \oplus PW_i')} \bmod P. \tag{1}$$

This is to see that the forged message C^* will pass the checking of Eq. (1) in the authentication phase.

$$\begin{aligned} C_2^*(C_1^{SK})^{-1} &= M^*(PW_i')^r((ID_i')^{r \cdot SK})^{-1} \\ &= M^*(PW_i')^r((PW_i')^r)^{-1} \\ &= M^* \\ &= (ID_i)^{t^*} \\ &= ID_i^{h(T^* \oplus PW_i')} \bmod P. \end{aligned}$$

Therefore, the system accepts the attacker's login request, making the Lin-Lai scheme insecure.

4 Proposed Scheme

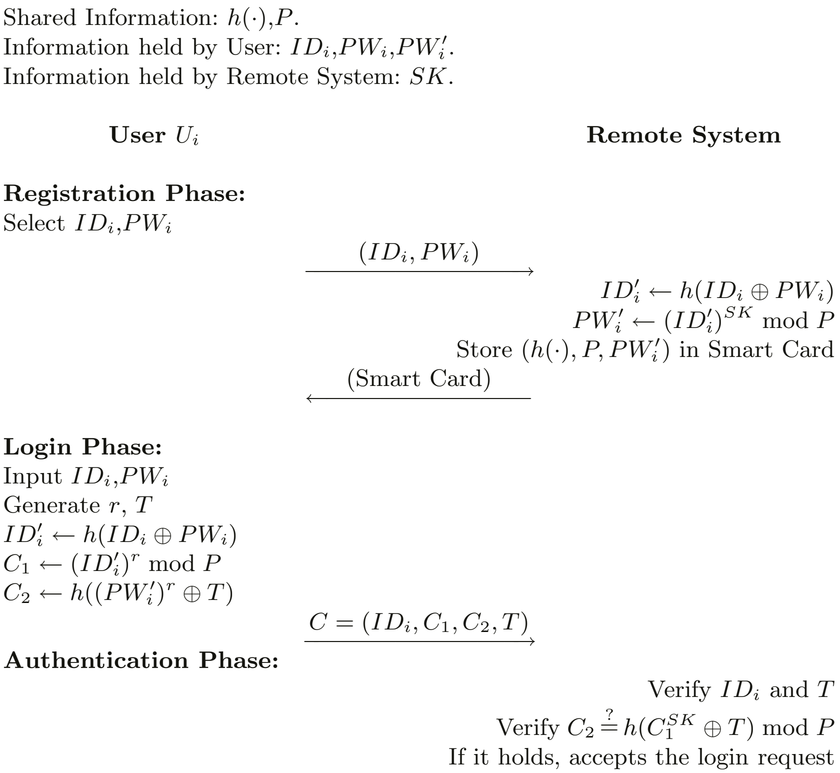
This section proposes an enhancement to the Lin-Lai scheme that can withstand impersonation attacks. The security of the proposed scheme is based on a one-way hash function and the discrete logarithm problem [8], and consists of a registration, login, and authentication phase. The registration phase in the proposed scheme is the same as in the Lin-Lai scheme. The proposed scheme is illustrated in Figure 1 and works as follows:

Login Phase: Whenever a user U_i wants to log-in, U_i has to insert his/her own smart card into the card reader and imprint the fingerprint. Then he/she types in identity ID_i and password PW_i. If U_i passes the fingerprint verification, U_i's smart card will perform the following operation:

1. Generate a random number r using the minutiae extracted from the imprint fingerprint.
2. Compute $ID_i' = h(ID_i \oplus PW_i)$.
3. Compute $C_1 = (ID_i')^r \bmod P$.
4. Compute $C_2 = h((PW_i')^r \oplus T)$, where T is the current timestamp of the input device..
5. Send a message $C = (ID_i, C_1, C_2, T)$ to the remote system.

Authentication Phase: After a transmission delay, the system receives the message C at time T', where T' is the receiving timestamp of the system. The system then performs the following operations:

1. The system checks whether the format of ID_i is correct or not. If the format is incorrect, the system rejects the login request.
2. If $|T - T'| \geq \Delta T$, where ΔT denotes the expected valid time interval for transmission delay, the system rejects the login request.
3. Verifying whether $C_2 \overset{?}{=} h(C_1^{SK} \oplus T) \bmod P$, if being successful, the system accepts the login request. Otherwise, the system rejects the login request.



Shared Information: $h(\cdot)$,P.
Information held by User: ID_i,PW_i,PW'_i.
Information held by Remote System: SK.

User U_i | **Remote System**

Registration Phase:
Select ID_i,PW_i

$\xrightarrow{(ID_i, PW_i)}$

$ID'_i \leftarrow h(ID_i \oplus PW_i)$
$PW'_i \leftarrow (ID'_i)^{SK} \bmod P$
Store $(h(\cdot), P, PW'_i)$ in Smart Card

$\xleftarrow{\text{(Smart Card)}}$

Login Phase:
Input ID_i,PW_i
Generate r, T
$ID'_i \leftarrow h(ID_i \oplus PW_i)$
$C_1 \leftarrow (ID'_i)^r \bmod P$
$C_2 \leftarrow h((PW'_i)^r \oplus T)$

$\xrightarrow{C = (ID_i, C_1, C_2, T)}$

Authentication Phase:

Verify ID_i and T
Verify $C_2 \overset{?}{=} h(C_1^{SK} \oplus T) \bmod P$
If it holds, accepts the login request

Fig. 1. Fingerprint-based remote user authentication scheme

5 Security Analysis

This section provides the proof of correctness of the proposed scheme. At first, we define the security terms [8] needed for security analysis of the proposed scheme as follows:

Definition 1. *A weak secret (user's password PW_i) is a value of low entropy $w(k)$, which can be guessed in polynomial time.*

Definition 2. *A strong secret (systems's secret key SK) is a value of high entropy $s(k)$, which cannot be guessed in polynomial time.*

Definition 3. *A secure one-way hash function $y = h(x)$ is one where given x to compute y is easy and given y to compute x is hard.*

Definition 4. *The discrete logarithm problem (DLP) is as follows: given a prime p, a generator g of Z_p^*, and an element $\beta \in Z_p^*$, it is hard to find the integer α, $0 \leq \alpha \leq p - 2$, such that $g^\alpha \equiv \beta \bmod p$.*

According to the above definitions, the following analyzes the security of the proposed scheme:

1. It is difficult for an attacker to derive the remote system's secret key SK from $PW_i' = (ID_i')^{SK} \bmod p$ because the complexity of computing SK from PW_i' is the discrete logarithm problem.
2. By the timestamp T, the remote system can check the correct time frame and prevent a replay attack. If an attacker modifies T into T^*, Step 2 in the authentication phase will fail.
3. No one can forge a valid $C = (ID_i, C_1, C_2, T)$. Because $C_1 = (h(ID_i \oplus PW_i))^r \bmod P$, an attacker has to get the PW_i in order to compute C_1. However, it is difficult for an attacker to get the PW_i, because the complexity of computing PW_i from C_1 is a one-way property of a secure one-way hash function and the discrete logarithm problem. Additionally, because $C_2 = h((PW_i')^r \oplus T)$, an attacker has to get the PW_i' in order to compute C_2. However, it is difficult for an attacker to get C_2 without knowing the user's password PW_i and the system's secret key SK. Therefore, the proposed scheme can resist impersonation attacks.

6 Efficiency Analysis

The computation costs of the Lin-Lai scheme and the proposed scheme in the registration, login, and authentication phases are summarized in the Table 1. The symbols in the Table 1 are as follows: T_{EXP} is the computation time for a modular exponentiation; T_{H} is the computation time for a one-way hash function; T_{M} is the computation time for a multiplication operation; and T_{XOR} is the computation time for an exclusive or operation.

In the registration, login, and authentication phases, the Lin-Lai scheme requires a total of 6 exponentiations, 4 hashes, 2 multiplications, and 4 bitwise exclusive or operations, but the proposed scheme requires only 4 exponentiations, 4 hashes, and 3 bitwise exclusive or operations. Therefore, the proposed scheme is more efficient than the Lin-Lai scheme.

Table 1. Comparisons of computation costs

	Lin-Lai scheme	Proposed scheme
Registration phase	$1T_{\mathrm{EXP}} + 1T_{\mathrm{H}} + 1T_{\mathrm{XOR}}$	$1T_{\mathrm{EXP}} + 1T_{\mathrm{H}} + 1T_{\mathrm{XOR}}$
Login phase	$3T_{\mathrm{EXP}} + 2T_{\mathrm{H}} + 1T_{M} + 2T_{\mathrm{XOR}}$	$2T_{\mathrm{EXP}} + 2T_{\mathrm{H}} + 1T_{\mathrm{XOR}}$
Authentication phase	$2T_{\mathrm{EXP}} + 1T_{\mathrm{H}} + 1T_{M} + 1T_{\mathrm{XOR}}$	$1T_{\mathrm{EXP}} + 1T_{\mathrm{H}} + 1T_{\mathrm{XOR}}$

7 Conclusion

The current paper demonstrated that an attacker can easily impersonate legal users to access the resources at a remote system in the Lin-Lai scheme. Thus, an enhancement to the Lin-Lai scheme was proposed that can withstand an impersonation attack. The proposed scheme removes the security flaw of the Lin-Lai scheme; moreover, it is more efficient than the Lin-Lai scheme.

Acknowledgements

This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessment).

References

1. Lamport, L.: Password Authentication With Insecure Communication. Communications of the ACM. Vol. 24. No. 11. (1981) 770-772
2. Lee, J.K., Ryu, S.R., Yoo, K.Y.: Fingerprint-Based Remote User Authentication Scheme Using Smart Cards. Electronics Letters. Vol. 38. No. 2. (2002) 554-555
3. ElGamal, T.: A Public-Key Cryptosystem And a Signature Scheme Based On Discrete Logarithms. IEEE Transactions on Information Theory. Vol. IT-31. No. 4. (1985) 469-472
4. Bae, I.G., Cho, B.H., Kim, J.S., Bae, J.H., Yoo, K.Y.: Online Fingerprint Verification System Using Direct Minutia Extraction. 13th International Conference on Computer Applications in Industry and Engineering. Honolulu, Hawaii. (2000) 120-123
5. Ratha, N.K., Karu, K., Chen, S., Jain, A.K.: A Real-time Matching System for Large Fingerprint Databases. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 18. No. 8. (1996) 799-813
6. Jain, A., Bolle, R., Pankanti S.: Biometrics Personal Identification In Networked Society. Kluwer Academic Publishing. (1999) 369-384
7. Lin, C.H., Lai, Y.Y.: A Fingerprint-Based User Authentication Scheme for Multimedia Systems. Proceedings of the 2004 IEEE International Conference on Multimedia & Expo (ICME 2004). Taipei, Taiwan. Vol. 2. (2004) 935-938
8. Menezes, A.J., Oorschot, P.C., Vanstone, S.A.: Handbook of Applied Cryptograph. CRC Press. New York. (1997)

A Study on the Correction of Gun Fire Error Using Neural Network

Yang Weon Lee[1] and Heau Jo Kang[2]

[1] Honam University, Seobong-Dong 59-1,
Kwangsan-Gu, Kwang-Ju, 506-714, South Korea
ywlee@honam.ac.kr
[2] Mokwon University, Doan-dong 800, Seo-gu, Daejeon, 302-729, South Korea
hjkang@mokwon.ac.kr

Abstract. Multilayer perceptrons trained with the backpropagation algorithm are derived for gun fire control system for miss distance correction and are compared to optimum linear filter based on minimum mean square error[1][2]. The structure of the proposed neural controller is described and performance results are shown.

1 Introduction

The dynamic control of aiming error by means of a derived feedback signal is at least a partial solution for a broadband spectrum of gunnery problem. Also, the principle of adjusting input controls on the basis of output performance (feedback control) is well known. Although it is a deceptively simple concept, new applications continue to occur, due not only to improvements in our understanding and modelling of physical systems, but also to the numerous analytical techniques developed in recent years. Many contributions have come in the area of stochastic optimal estimation and neural approach: techniques enabling the extraction of useful information from error-corrupted data, and the subsequent improved control over many systems subject to random disturbances.

The problem to be addressed involves the on-line derivation of gun fire control adjustments to minimize the measured miss distance between a target and the projectiles from the naval gun. The class of errors to be minimized are those which cannot be removed either in the gun design phase or by precalibration exercises.

Gun process errors in operational systems arise from many sources. Examples are ship's motion, cyclic meteorological conditions, gun jump, aerodynamic variability among projectiles, and rising gun temperature and bore erosion during a continuous firing sequence.

It is well known that above mentioned errors exist, but little has been done to remove them, primarily since they are a fairly complex stochastic process, and the procedures for extracting and using information from such processes have not, as yet, been applied [1][2]. In the following, the mathematical model of the

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 339–345, 2005.

error process will be presented, the neural control law for such a system developed, and numerical examples presented to compare the performance between optimal control and neural control quantitatively.

2 System Concept

The gun system is comprised of the target tracking radar, the gun fire control system, and the gun itself. By adding a miss distance measuring capability - which is practical - the feedback control concept that is to be considered is shown in Fig. 1. The modified system incorporates prediction of gun error and includes this prediction in the gun orders from the fire control system. It is noted that the radar, fire control, and gun constitute an open loop fire control system; the new components provide closed loop gun fire control of gun error, but the total system remains an open loop control system with respect to the basic gun fire control operation.

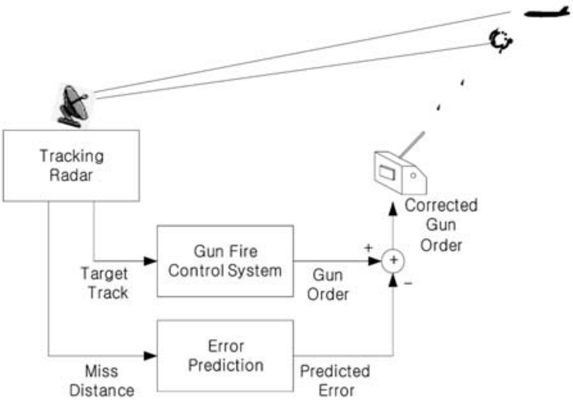


Fig. 1. The closed loop gun fire control system using the measured miss distance

3 Optimal Correction Solution

3.1 Gun Process

A mathematical model of gun system performance is a vector valued discrete random process for which the components of the vector represent the different components of performance, and the successive discrete values of time correspond to the successive projectiles in a firing sequence. The gun process $\{X_n; n = 1, 2, \ldots\}$ can be represented as the sum of two processes: an aiming process that will be denoted α_n; and a ballistic process that will be denoted β_n. Each will be a discrete random process, giving

$$X_n = \alpha_n + \beta_n. \tag{1}$$

The α-process will in general be a correlated process and could represent such systematic and time dependent effects as gun temperature and wear, air density and wind variability, and all such phenomena that ideal representation and ideal processing by a perfect fire control system could nullify. The α- process can be modelled as a linear combination of the previous variate and a new random variable, θ, that is orthogonal to all preceding variates, as follows,

$$\alpha_n = C_n \alpha_n - 1 + K_n \theta_n, \tag{2}$$

where C_n and K_n are scalars that could depend on n, and θ_n is the nth random variable in the sequence $\{\theta_n; n = 1, 2, \ldots\}$, specified by

$$\begin{cases} E\{\theta_n\} = 0, n = 1, 2, \ldots \\ E\{\theta_n \theta_m\} = \begin{cases} \sigma_\theta^2 & ; n = m \\ 0 & ; n \neq m \end{cases} \end{cases} \tag{3}$$

The β-process has no correlated component and represents performance variability among the individual projectiles. The mean value $E\{.\}$ is specified by

$$\begin{cases} E\{\beta_i\} = 0, i = 1, 2, \ldots \\ E\{\beta_i \beta_j\} = \begin{cases} \sigma_\beta^2 & ; i = j \\ 0 & ; i \neq j \end{cases} \end{cases} \tag{4}$$

Note that the variance is independent of i (that is, time) and that no assumption is made regarding either the density or distribution functions for the β-process.

3.2 Mathematical Description, *X*-Process

The required description of the typical channel, one of the three components in the projectile space, can be described by a scalar random process as following :

$$x_n \equiv \alpha_n + \beta_n + \gamma_n + u_n, \tag{5}$$

where α_n, β_n, γ_n are prediction, aiming and ballistic errors at time n respectively and u_n is the closed-loop error-correction term. The closed-loop error corrections are determined from measurements of the miss between the intended impact points and actual impact points. Referring to (5), the miss distance x_n is measured, and the correction u_n to the gun orders for the shell to be fired next is computed. The model for the gunnery errors is represented in this way :

$$d_k = y_k - u_k, \tag{6}$$

$$y_k = x_k + \eta_k, \tag{7}$$

and

$$\begin{aligned} x_n &= s_n^{'} - s_n, \\ &= d_n + u_n. \end{aligned} \tag{8}$$

where s_n and s'_n denote the target and the spotting position at time n ,η_k represents the measurement error and $d_n \triangleq \alpha_n + \beta_n + \gamma_n$. The closed-loop error correction system configuration is shown in Fig. 2.

The purpose of the error correction filter is to ensure $E[x_n] \ll E[d_n]$ and $var[x_n] \ll var[d_n]$. Corn[2] assumed that the aiming error is a second order Markov process, and the ballistic error is a white gaussian noise with $N(0, \sigma_b^2)$. Then the mean and variance of the errors for Corn's correction predictor[2] become

$$E[x_n] = 0, \tag{9}$$

$$var[x_n] = \sigma_\theta^2 + \sigma_b^2 - \rho^{2p}(\sigma_\theta^2 - V_{n-p})^2, \tag{10}$$

$$V_n = \frac{\sigma_\theta^2(1-\rho^2) + \rho^2 V_{n-1}}{\sigma_\theta^2(1-\rho^2) + \sigma_b^2 + \rho^2 V_{n-1}} \sigma_b^2, \qquad \forall n \geq 2, \tag{11}$$

where ρ and V_{n-p} are the correlation coefficient of x_n and the Kalman filter variance in estimating α_{n-p} respectively. We can see the effect of the error predictor from the variance of d_n,

$$var[d_n] = \sigma_\theta^2 + \sigma_b^2. \tag{12}$$

However, in general, the gun error d_n cannot be assumed as a linear combination like (5). Here we propose a new CLFC system based on a neural network controller, which is more robust over the non-linear, high-order system[3].

However, in general, the gun error d_n cannot be assumed as a linear combination like (5). Here we propose a new CLFC system based on a neural network controller, which is more robust over the external disturbances.

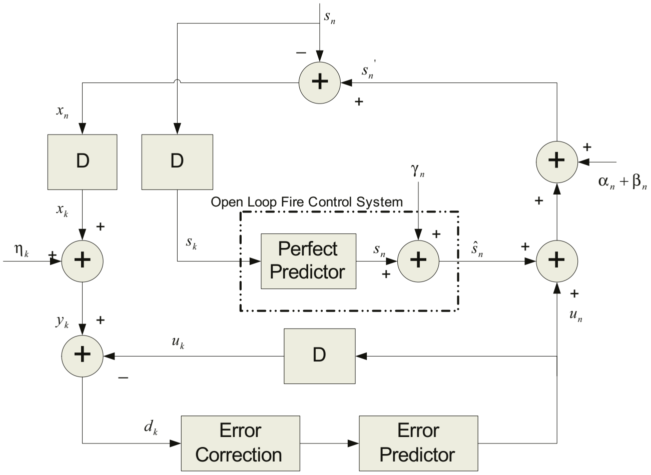


Fig. 2. An error model of the CLFC system

4 Neural Solution

As shown in Fig. 3 the architecture for the proposed neural network controller consists of two Multi-Layer Perceptrons(MLP). The first network predicts miss distance at time step k by using the past data from d_{k-1} to d_{k-4}, and the second network predicts the miss distance at time step n by using the previous predictor's output and past data from d_{k-1} to d_{k-4}. Here, we can see that the controller is a high order system since the first and the second networks have delayed inputs. In case of Corn[2], error prediction at time n is calculated by linear transformation at time k . But the neural network not only uses all possible past data but also does not effect the nonlinear properties of gun error processes.

The operation of this neural controller has two modes: learning and working modes. When the controller is in the learning mode, output error is back propagated through the network and internal weights are updated. And in the real working mode, the network produces the output. The network shown in Fig. 3 will be easily expanded to the three dimensions for representing the projectile in real space.

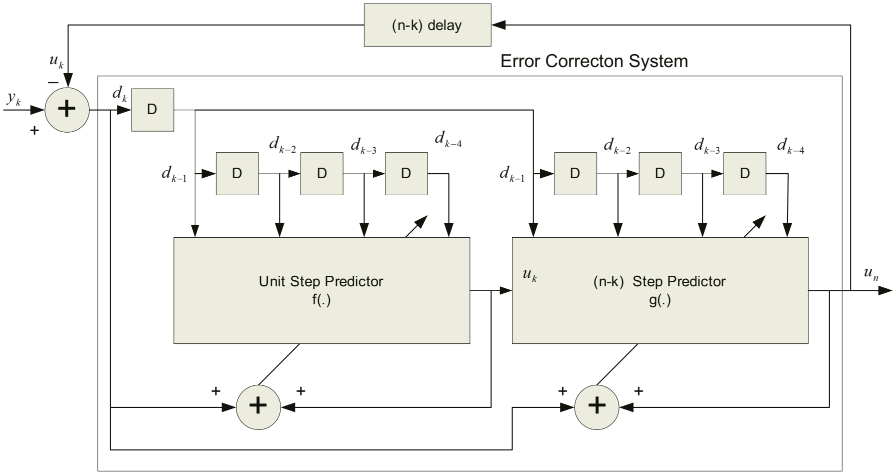


Fig. 3. Error correction system using MLPs

5 Performance Evaluations

The feasibility of the proposed neural controller to correct the miss distance is evaluated by the computer simulation. The considered target is at a constant speed(300m/s) and nonmaneuvering, and flies directly at the gun. The 30 mm gun is represented by a constant rate of fire(600 RPM) and an error model consisting of a range-dependent bias, a serially correlated error, and an uncorrelated error in azimuth and elevation, with azimuth and elevation components of errors assumed independent. The projectile is modelled by an exponential fit to range-table data of the 30 mm gun:

$$T = C(e^{KR} - 1). \tag{13}$$

where T is the time of intercept, R is the range, and C and K are constants for the projectile. And the projectile velocity is used by 900 m/s. The maximum and minimum intercept ranges are 4,000 m and 500 m each. The number of rounds to be fired is determined from these intercept ranges, the gun's firing rate, and projectile's and target's dynamics. Firing will start when the first round is to intercept the target at maximum intercept range and end when the last round is to do so at minimum intercept range. The proposed neural network consist of input and output layer and two hidden layers whose layer has 20 nodes each. Also, error back-propagation algorithm is used for training and the number of training is two times. And the time of unit delay D is calculated by gun's firing rate. Fig. 4 shows the total rms miss distance as a function of rounds in meter scale for each algorithms, and Table 1 shows the mean and the variance. As shown in the experimental results, our neural controller for error correction system is superior to Corn's filter[2] and the Kalman predictor[1]. Thus, it is clear that the neural approach for CLFC holds considerable promise for achieving accurate gun fire control and, furthermore, the technology for implementing these control policies is available.

Table 1. Comparison of the mean and the variance

Filter Type	Mean $[m]$	Variance $[m]$
OLFC	19.93	11.91
CLFC with Neural	3.03	3.64
CLFC with Kalman	7.49	8.39

6 Conclusions

In this paper, we propose a neural network controller for closed loop gun fire control system. In addition to the derivation, the potential improvement to be expected from the application of such configuration to gun fire control systems has been evaluated. The simulation results show that the rms errors in mean and variance can be reduced significantly to around 3 meters. Also, the specification of a hardware design indicates that such a controller can be implemented on either a new or existing system for a relatively small investment.

Although the error processes analyzed in this paper are for gun fire control system, it is expected that the elements of neural scheme presented will find application to a wide range of subjects.

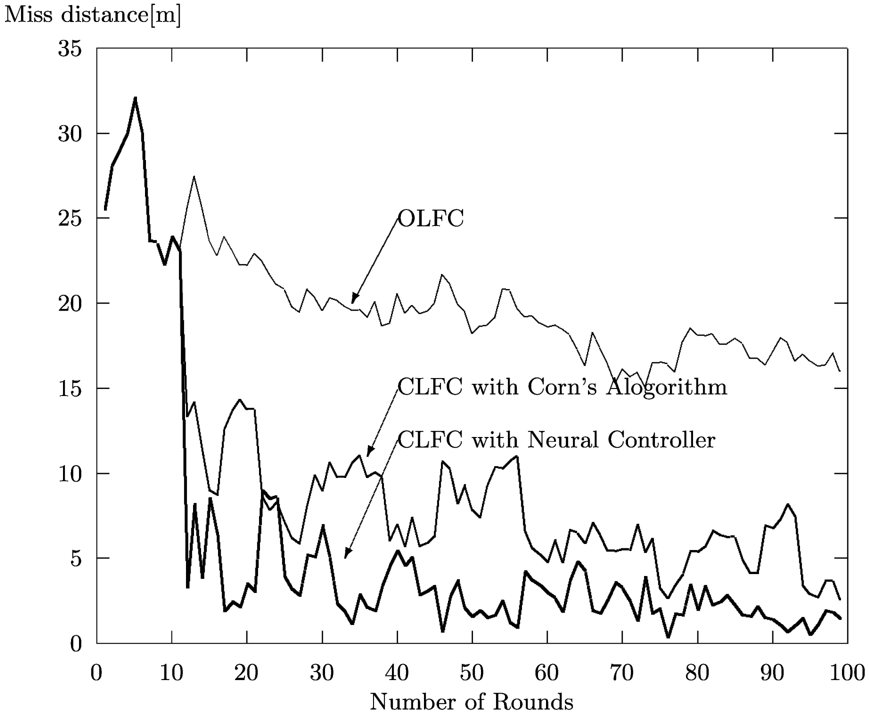


Fig. 4. Total miss distance in rms value [m]

References

1. R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, Vol. 82D, Mar. 1960, pp.35-44
2. R. J. Corn, "An analysis of closed loop control of gun systems," *System Evaluation Group, Center of Naval Analysis*, Nov. 1971.
3. K. S. Narendra and K. Parthasarathy, "Identification and control of dynamical systems using neural networks," *IEEE Trans. on Neural Networks* ,1990,pp.4-21

An Efficient Moving Object Extraction Algorithm for Video Surveillance

Da-Jinn Wang[1], Thou-Ho Chen[2], Yung-Chuen Chiou[2], and Hung-Shiuan Liau[2]

[1] Department of Information Management,
National Kaohsiung Marine University, Kaohsiung, Taiwan
Wangdaj@mail.nkmu.edu.tw
[2] Department of Electronic Engineering,
National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan
thouho@cc.kuas.edu.tw, happyout@ms14.hinet.net,
1093320124@cc.kuas.edu.tw

Abstract. In this paper, an efficient moving object extraction algorithm for surveillance application is proposed which employ change detection strategy to obtain motion information of moving-object instead of complex operator. In addition, background subtraction is introduced to solve the problems of still object and uncovered background which is generally ill-inherency existed in conventional method. After that, the internal part region of moving-object may be confused with real-static region due to frame difference used. Hence, we use the concept of region adjacent graphic to overcome it. Finally, a post-processing step is used to remove noise regions and refine the shape of objects segmented. Moreover shadow effects can be suppressed in the pre-processing step. Experimental results demonstrate various results of segmented video sequence for both indoor and outdoor scenes and show that the proposed algorithm is superior to others in terms of obviating static region of internal part of moving-object and edge defects.

1 Introduction

Conventional video-coding standards (e.g., H.261/3 or MPEG-1/2) that employ the frame-by-frame coding policy, can't provide high-level feature of video contents. The MPEG-4 coding standard [1] has introduced the concept of a video-object plane (VOP) as the basic coding element for supporting visual multimedia communication and will be applied to many multimedia content descriptions. Each VOP contains the shape and texture information of semantically meaningful object in the scene. In order to encode video objects using object-by-object in video sequences rather than frame-by-frame and achieve the content-based manipulation for video content, automatic video segmentation will play an important role of deriving VOP from video sequences in MPEG-4 video part.

2 The Proposed Segmentation Algorithm

The block diagram of the proposed video segmentation algorithm is described in Figure 1, which is indicated to separate the moving-object regions from other parts of the

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 346–352, 2005.

scene by using motion information. We construct and maintain an updating background reference from the static map, unlike some reported algorithms which use the shape features of moving-object regions. The details of each module in the proposed algorithm, will be discussed in the following subsections

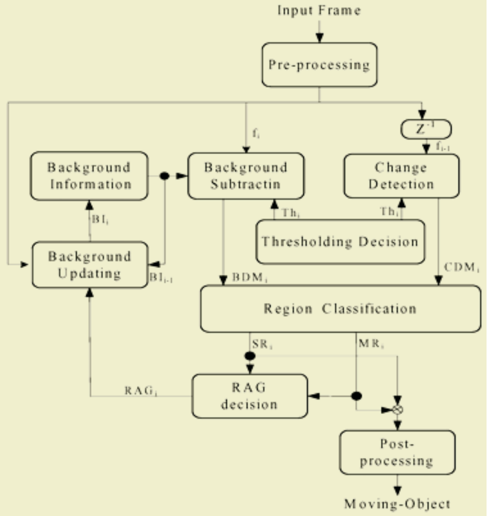


Fig. 1. Block diagram of proposed segmentation algorithm

2.1 Frame Difference

Frame difference technique [2][3] is often used in change detection based segmentation algorithm, that give difference between two consecutive input frames. However, since the behavior and characteristics of moving objects differ manifestly in real-situation, the quality of threshold frame difference depends strongly on environmental noise, light change and object motion. Therefore, stable and concerting object information is very hard to acquire. In addition, if the threshold difference mask can not be obtained automatically, these kinds of video object segmentation algorithms are realized hard in real systems. A robust method of threshold decision which is proposed in next subsection.

2.2 Threshold Decision

In this section, a way that can class the correspondent pixel as either foreground class or background pixel. The proposed block diagram of frame difference is shown in Figure 2. It includes threes parts: *Histogram Analysis, Statistics Analysis, Pixel Classification.*

Histogram Analysis

The histogram is constructed from frame difference, and provides information of the graylevel distribution to analyze the characteristic itself. Hence, we find the gray bin of maximum number of histogram almost related to background information to estimation parameter of background part, denoted, $His(d_k)$ and d_k is pixel of difference image located on k position.

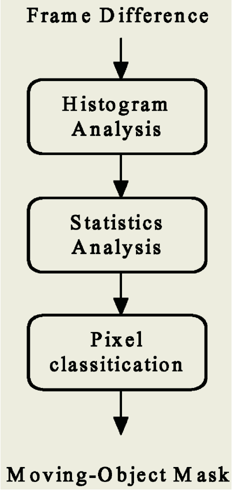


Fig. 2. Block diagram of generating a moving-object mask

Statistic Analysis

Therefore, the mean and variance of the pixel which the position is located at the gray-level the same with gray bin of maximum number in histogram distribution should be considered, then the mean and variance can is calculated within window mask, as follow:

$$\mu_{w_i}(d_k) = \frac{1}{N} \sum_{j=1}^{N} w_i(j) \tag{1}$$

$$std_{w_i}(d_k) = \sqrt{\frac{1}{N} \sum_{j=1}^{N} \left[w_i(j) - \mu_{w_i}(d_k) \right]^2} \quad for\ i = 1, 2, \ldots, His(d_k) \tag{2}$$

where w_i means window size, and its size is N.

Pixel Classification

Last, the pixel in the difference image will be classified to foreground parts or background parts according to following formula. Let μ_b and std_b present the estimated parameter of background in current frame difference, and the sign, c is constant, dependent on video content.

$$\begin{array}{ll} \mu_b = \dfrac{1}{his(d_k)} \displaystyle\sum_{i=1}^{his(d_k)} \mu_{w_i}(d_k) & f\left(\mid d_k - \mu_b \mid > c * std_b \right) \\ & \quad foreground\ pixel \\ std_b = \dfrac{1}{his(d_k)} \displaystyle\sum_{i=1}^{his(d_k)} std_{w_i}(d_k) & else \\ & \quad background\ pixel \end{array} \tag{3}$$

Based on previous threshold decision, threshold curve be calculated on sequences of Hall Monitor and Lerry shown in Figure 3, and results of pixel classification shown in Figure 4.

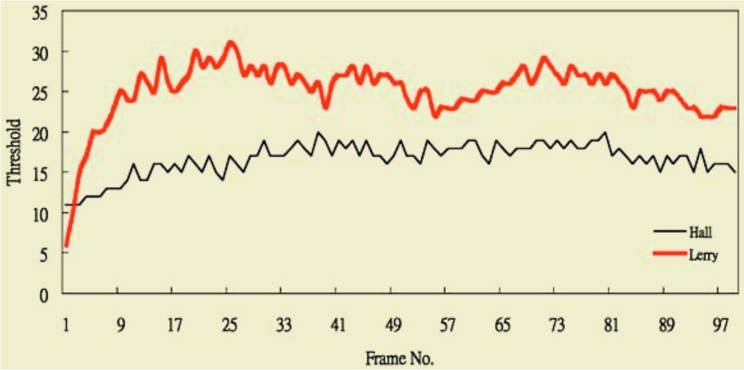
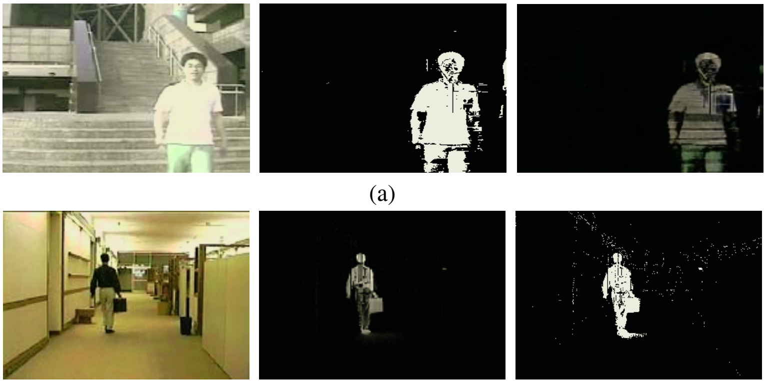


Fig. 3. Threshold curve for Hall Monitor and Lerry sequence


(a)

(b)

Fig. 4. Result of pixel classification from left to right, original frame, frame difference image, and binary image (a) Lerry sequence (c =10); (b) Hall Monitor sequence (c = 10)

2.3 Region Classification

Region Classification is used to classify the region either moving-region or nonmoving-region by combing frame difference mask (FDM) and background subtraction mask (BDM). Assume, set MR is corresponding to moving object and SZ is corresponding to static zone in i*th* frame after region classification decision is applied. Moreover, they can be formularize following.

$$MR = \left\{ AP_i \middle| fdm_i \cap bdm_j = 1 \, \forall fdm_i \in FDM, \forall bdm \in BDM \right\} \tag{4}$$

$$SZ = \left\{ AP_i \middle| fdm_i \cap bdm_j = 0 \, \forall fdm_i \in FDM, \forall bdm \in BDM \right\} \tag{5}$$

where AP_i denote “active pixel” in pixel i index. Both MR and SZ regions can be shown in Figure 5.

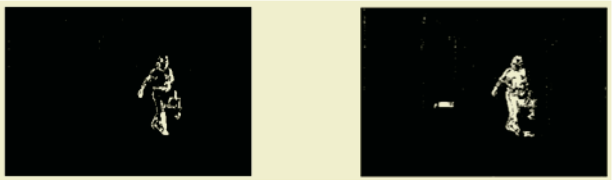


Fig. 5. Results of pixel classification. (a) MR region; (b) SZ region

2.4 Region Adjacent Graphic

Our machination of background updating is achieved which the static region (i.e., SZ region) is treat as input. In general, the static zone may contain three types: real static-region, light change, and static part region of moving-object. Actually, we hope the static map that it only involves real-static object and light change, but it is unlikely due to region classification is used. In order to let static part region of moving-object is not regarded as real static-object to update the background image; therefore, region adjacent graphic (RAG) [4] can be solve this problem. We can find a fact that if a region is not real static-object, it must belong to a part of moving-object and connected to moving region (i.e., MR region). Hence, it can be utilized to decide the region whether real static-region or not. In Figure 6 which illustrate the situation.

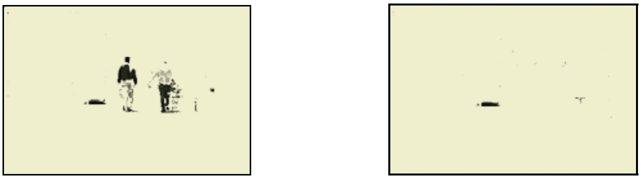


Fig. 6. Non-static region effect. (a) Static part region of moving-object before using RAG decision(b) Static part region of moving-object is removed

2.5 Background Updating

After RAG decision is done, background updating [5][6] will begin. A stationary map (SM) is need to store if the current pixel belongs to background. The background updating can be shown as following equation:

$$SM_i(x,y) = \begin{cases} SM_{i-1}(x,y)+1 & if\ (x,y)\ is\ inactive\ pixel \\ 0 & else \end{cases}$$
$$BI_i(x,y) = \begin{cases} f_i(x,y) & if\ SM_i(x,y) = F_{th} \\ BI_{i-1}(x,y) & otherwise \end{cases} \tag{6}$$

Where $SM_i(x,y)$ and $BI_i(x,y)$ are stationary map and background buffer at i_{th} frame; $f_i(x,y)$ is current frame; F_{th} is stationary threshold. In addition, "inactive pixel" means unchanged in result of RAG decision. Figure 7 illustrated this process, subimage (a) is original frame of Hall Monitor sequence, and (b) is correspond to background image of Hall Monitor sequence in 49 frame.

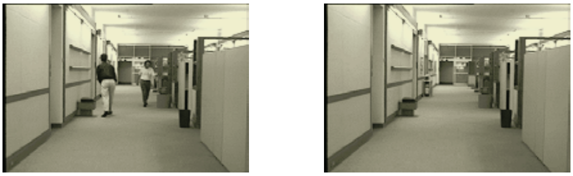


Fig. 7. Background updating image. (a) Hall Monitor frame 49 ；(b) Updated background image at frame 49

2.6 Pre-processing and Post-processing

In pre-processing step, it contains two principle operators, Gaussian smoothing filter and Sobel operator [7]. In former operator, the goal is to reduce the noise effect of input frame and the Sobel operator is applied on input video frame to deal with shadow effect, it is shown in figure 8. In post-processing, the connected components algorithm [8] is used to eliminate small region, and boundary is refined by open-close operator.

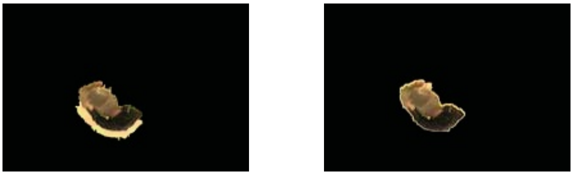


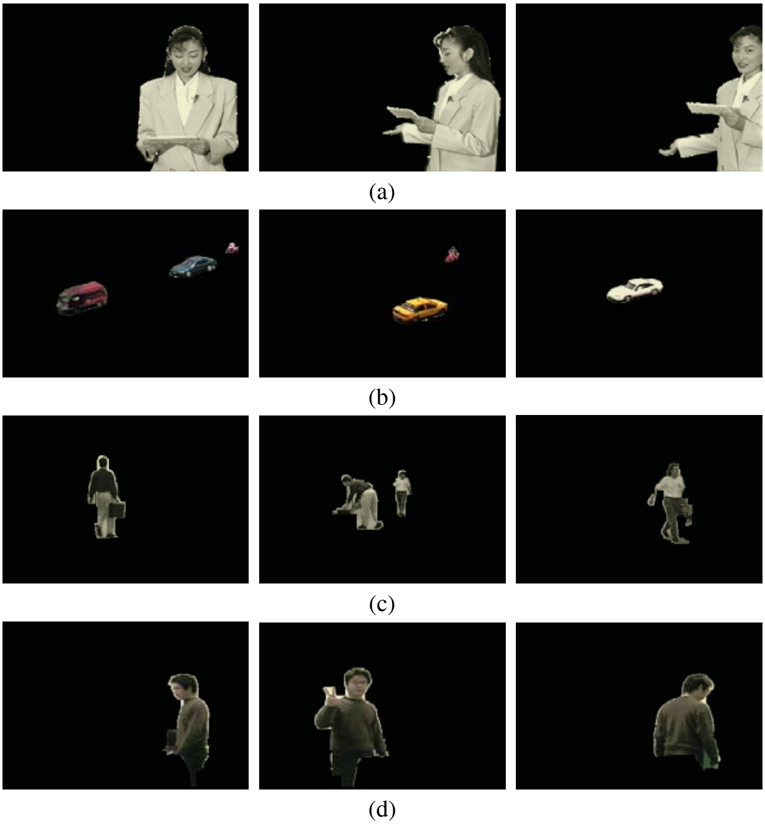
Fig. 8. Removing of shadow effect: (a) segmentation result with shadow effect; (b) segmentation result after a sobel operator applied

3 Experimental Results

The performance of the proposed video segmentation algorithm is simulated with several video sequences, *Weather*, *Car, Hall Monitor*, plane, *Lerry* shown in Fig. 9. The quality of segmentation results is evaluated subjectively. The segmentation results of sequence In subimage (a) and (b), even if the foreground object move a lot, they can also be correctly segmented. In addition, subimage (c) show that the shadow effect and light changing occur in both sequences can be removed.

4 Conclusion

In this paper, an efficient video segmentation algorithm for surveillance is proposed. The change detection and background updating method is utilized to obtain the motion information and construct reliable background information. Then, region classification can be produced by combining the FDM with BDM. After that, region adjacent graphic is applied on static map to exclude non-real static region. In the further, the pre-pressing can be reduce shadow effect, and post-processing is applied to remove noise and refine boundary. Finally, the experimental results demonstrate that good segmentation quality can be obtained efficiently.



(a)

(b)

(c)

(d)

Fig. 9. Segmentation results of four sequences: (a) Weather, (b) Car, (c) Hall Monitor, (d) Lerry

References

1. T. Sikora,“The MPEG-4 video standard verification model,”*IEEE Trans. on Circuits Syst. Video Technol*., vol. 7, pp. 19-31, Feb. 1997.
2. T. Aach, A. Kaup and R. Mester, “Statisticl model-based change detection in moving video,” *Signal Processing*, vol. 31, pp.165-180, Mar. 1993.
3. Jinhui Pan; Chia-Wen Lin; Chuang Gu; Ming-Ting Sun, “A robust video object segmentation scheme with prestored background information,” Circuits and Systems, 2002. *ISCAS 2002. IEEE International Symposium* on Volume 3, 26-29 May 2002 Page(s):803 - 806
4. R.Jain, R.Kasturi, and B.G. Schunck, *Machine Vision*. Reading, McGraw-Hill, 1995.
5. Alice Caplier, Laurent Bonnaud, Jean-Marc Chassery, “Robust Fast Extraction of Video Object s Combining Frame Differences and Adaptive Frame Reference Image,” *IEEE International Conference on Image Processing*, vol. 2, pp.785-788, 2001.
6. Andrea Cavallaro and Touradj Ebrahimi, “Video object extraction based on adaptive background and statistical change detection,” *Proc. of IEEE Visual Communications and Image Processing*, pp.465-475, January 2001.
7. R.C. Gonzalez and R. E. Woods, *Digital Image Processing*. 2ED, Prentice-Hall Inc., 2002.
8. R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision*. Reading, MA: Addison-Wesley, 1992, pp. 28-48.

An Experimental Comparison on Gabor Wavelet and Wavelet Frame Based Features for Image Retrieval

Yu-Long Qiao[1], Jeng-Shyang Pan[1,2], and Sheng-He Sun[1]

[1] P.O. Box 339, Department of Automatic Test and Control,
Harbin Institute of Technology, Harbin, Heilongjiang, China
qiaoyulong@dsp.hit.edu.cn
[2] Department of Electronic Engineering,
National Kaohsiung University of Applied Sciences, Taiwan

Abstract. The mean and standard deviation of the magnitudes of coefficients of the transform domain are common used in texture image retrieval. This paper performs an experimental comparison on Gabor wavelet and wavelet frame based features for this application. The detail comparison results indicate the performance of the former is not superior over that of the last in the query-by-example retrieval.

1 Introduction

Digital image libraries are becoming larger and more visual information is put in digital form as well as on-line. To make use of those sources, there must be an effective method to manage and search those images. The drawbacks of the text-based database management system and image retrieval system become more and more obvious. Thus, content-based image retrieval system has drawn large attention of researcher. Texture, an important feature for image retrieval, is considered in this paper.

W. Y. Ma and B. S. Manjunath [1] used orthogonal (16-tap Daubechies wavelet) and bi-orthogonal (5/3 bi-orthogonal wavelets) wavelet transform and Gabor wavelet transform to extract features and got that Gabor wavelet is the best one for image retrieval. So in their paper [2], they proposed 48 Gabor wavelet features for browsing and retrieval of image data. M. kokare et. al [3] proposed cosine-modulated wavelet based texture feature for image retrieval and showed that it is better than Gabor wavelet based feature.

Although the discrete wavelet transform possesses many usefully characteristics. It is translation-variant, that is to say, when a texture pattern is translated, its numerical descriptors will be modified but not translated, which is not expectative in texture analysis. So Unser [4] proposed discrete wavelet frame representation for texture. In this paper, we perform an experimental comparison on Gabor transform and wavelet frame based features for image retrieval. The experimental results indicate they have similar performance on this application.

2 Basic Theories

2.1 Gabor Transform Based Features

Ma and Manjunath [2] applied Gabor filters on a texture and generated 24 channels. Mean and standard deviation of the magnitudes of coefficients of each channel

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 353–358, 2005.

$$\mu = \frac{1}{N^2}\sum_{i,j=1}^{N}\left|Tf(i,j)\right| \text{, } \sigma = \sqrt{\frac{1}{N^2-1}\sum_{i,j=1}^{N}\left(\left|Tf(i,j)\right|-\mu\right)^2} \qquad (1)$$

are computed to form the feature vector for image retrieval. $Tf(\cdot,\cdot)$ denotes the transformed coefficient. Their experiments show this feature vector is the best one. We noticed that they used wavelet transform and tree-structured wavelet transform. Those transforms are complete, while Gabor transform is over-complete. Unser [4] explained the discrete wavelet frame is better than the discrete wavelet for representing texture. So we will compare Gabor wavelet and wavelet frame based features for image retrieval.

2.2 Discrete Wavelet Frame Transform (DWFT)

Since Mallat [5] presented the wavelet representation and some applications, the wavelet has been successfully applied in various fields of signal processing. Although the discrete wavelet transform possesses many usefully characteristics. It is translation-variant, that is to say, when a texture pattern is translated, its numerical descriptors will be modified but not translated, which is not expectative in texture analysis. So Unser [4] proposed discrete wavelet frame representation for texture.

The implementation of the discrete wavelet frame transform is similar to that of the discrete wavelet transform, except that there is no downsampling operation. For a one-dimensional signal *x*, the block diagram of DWFT is shown in Fig.1.

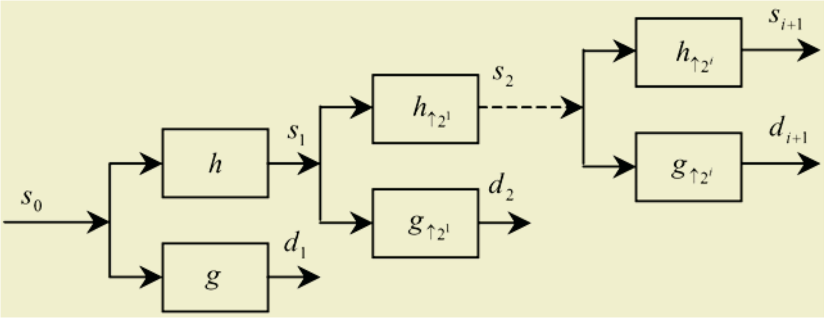


Fig. 1. The block diagram of DWFT

This algorithm is

$$\begin{aligned} s_{i+1}(k) &= h_{\uparrow 2^i} * s_i(k) \\ d_{i+1}(k) &= g_{\uparrow 2^i} * s_i(k) \end{aligned} \qquad (2)$$

with s_0=*x*. It involves two basic filters *h* and *g*. At each step, a branch output of the last step, s_i, convolutes with the filters $h_{\uparrow 2^i}$ or $g_{\uparrow 2^i}$, where $[\cdot]_{\uparrow 2^i}$ denotes the upsampling by a factor of 2^i, which also means inserting 2^i-1 zeros between two taps of the basic filters. The decomposition of an image can be obtained by successively one-dimensionally processing along the rows followed by the columns. The linear Battle-Lemarié spline wavelet (BLS), “least symmetric” compactly supported wavelets corresponding to 8 (DL8) and 10 (DL10) taps low-pass filter, 9/7 bi-orthogonal wavelets (B97) will be employed in our paper.

3 Experiments and Results

The image database consists of 112 different brodatz textures [7]. Every texture is size of 512×512 and is divided 16 128×128 non-overlapping subimages, thus there are 1792 texture images. For every texture pattern, there are another 15 patterns belonging to the same class. We take a texture pattern from every class as the query texture. Another 15 patterns are stored in the database for retrieval. Thus, there are 1680 texture images in the database. After an experiment, these 112 query texture patterns are replaced by another 112 texture patterns from different classes such that every texture pattern can serves as the query texture in the retrieval experiment.

The normalized city-block distance

$$d(\mathbf{q},\mathbf{f}_i) = \sum_m \left|q^m - f_i^m\right| \Big/ \sigma(f^m) \tag{3}$$

is used as similarity measure of different textures. **q** and $\mathbf{f}_\mathrm{i}$ are the feature vectors of the query texture pattern and the *i*th pattern in the database. $\sigma(f^m)$ is the standard deviation of the respective feature over the entire database. We adopt top *K* matches method in the experiment, that is, the distances between the query texture pattern and the texture patterns in the database are computed and arranged in ascending order, the *K* closest texture patterns are the retrieval result. The average retrieval rate (ARR) of the top *K* matches method is defined as the average percentage number of texture pattern belonging to the same class as the query pattern in the top *K* matches. It measures the retrieval performance.

In the first experiment, we perform five-level wavelet frame decomposition on each texture and obtain 16 subbands, then compute mean and standard deviation of the magnitudes of coefficients of each subband except the lowest frequency subband to form a 30-dimensional feature vector. When the top match numbers are 15 and 100, the ARRs of the employed wavelet frames and Gabor wavelet [2] (Gabor) are listed in Table 1 and Table 2. It can be seen that the performances of the wavelet frames based features are little worse than that of Gabor wavelet based features. The minimal difference is 1.12% when the top match is 15. In table 2, the maximal difference is 0.9%. Fig.2 shows the results for top matches from 15 to 100. It also indicates the little difference among these wavelets.

Table 1. ARRs for top match 15

Wavelet	Gabor	DL8	DL10	BLS	B97
ARR(%)	74.71	73.59	73.50	73.19	73.23

Table 2. ARRs for top match 100

Wavelet	Gabor	DL8	DL10	BLS	B97
ARR(%)	93.07	92.17	92.11	92.25	92.31

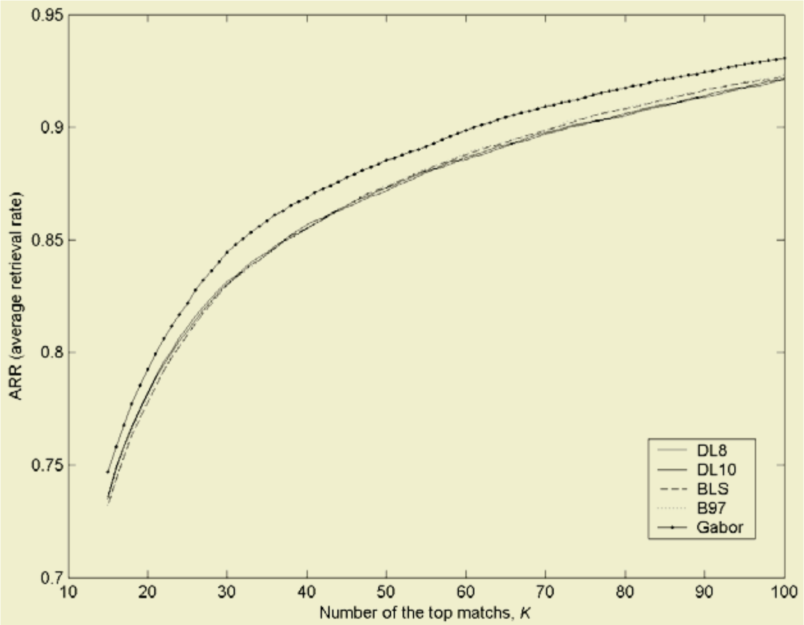


Fig. 2. ARRs vs. number of top matches

The second experiment performs four-level wavelet frame decomposition on each texture and features contain mean and standard deviation of the magnitudes of coefficients of the lowest frequency subband, so there are 26 features. For comparison, the Gabor wavelet based features include mean and standard deviation of the texture. When the numbers of top matches are 15 and 100, the ARRs of the employed wavelet frame and Gabor wavelet are listed in Table 3 and Table 4. When the number of top matches is 15, ARR of Gabor wavelet is worse than that of other wavelet frame. In Table 4, the performance of Gabor wavelet is better. All performance differences between Gabor wavelet and other wavelet frames of the two tables are small. The maximal difference in table 3 is 0.46, which corresponds to BLS. Figure 3 shows ARRs vs. number of top matches, which further indicates their performances are similar.

Table 3. ARRs for top match 15 when the lowest frequency subband is included

Wavelet	Gabor	DL8	DL10	BLS	B97
ARR(%)	76.83	77.12	77.04	77.29	77.22

Table 4. ARRs for top match 100 when the lowest frequency subband is included

Wavelet	Gabor	DL8	DL10	BLS	B97
ARR(%)	93.68	93.32	93.31	93.50	93.47

The results indicate that the performances of all the employed wavelet frames based features almost equal to that of Gabor wavelet. In the experiments, we notice that the Gabor wavelet based feature extraction time is higher than that of other wavelet frames. At the same time, the dimension of the feature vector based on the wavelet frame is smaller than that of Gabor wavelet so that it will save much retrieval time. In consequence, it will properly reduce computational complexity if one of the employed wavelet frame is adopted.

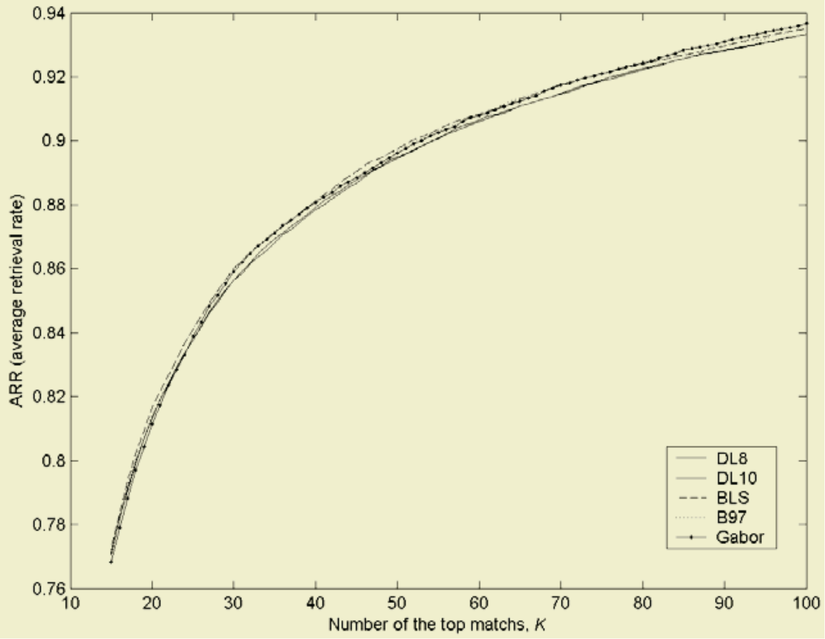


Fig. 3. ARRs vs. number of top matches of all wavelet frames when the lowest frequency subband is included (Gabor wavelet based features include the mean and standard deviation of the texture)

4 Conclusion

This paper performs a comparison between Gabor wavelet based features and wavelet frames based on features. The results indicate the Gabor wavelet is not superior over the wavelet frame on the query-by-example retrieval. The wavelet frame based features have advantages on feature extraction time, retrieval performance and searching time. However, it does not mean the Gabor wavelet can be replaced with the wavelet frame. Because the Gabor wavelet has many properties that the wavelet frame may not possess: adaptive 4 filters for retrieval [2], characterizing perceptual attributes [8] etc.

References

1. Ma W. Y., Manjunath B. S.: A Comparison of Wavelet Transform Feature for Texture Image Annotation. Proc. ICIP, 2 (1995) 256-259

2. Ma W. Y., Manjunath B. S.: Texture Feature for Browsing and Retrieval of Image Data. IEEE Trans. Pattern Anal. Machine Intell. 18(8) (1996) 837-842
3. Kokare M., Chatterji B. N., Biswas P. K.: Cosine-modulated Wavelet Based Texture Features for Content-based Image Retrieval. Pattern Recognition Letter, 25 (2004) 391-398
4. Unser M.: Texture Classification and Segmentation Using Wavelet Frames. IEEE T. IP. 4(11) (1995) 1549-1560
5. Mallat, S. G.: A Theory for Multiresolution Signal Decomposition: the Wavelet Representation. IEEE Trans. Pattern Anal. Machine Intell. 11(7) (1989) 674-693
6. Daubechies I.: Ten Lectures on Wavelet. Philadephia PA: SIAM, (1992)
7. Brodatz P.: Textures: A Photographic Album for Artists & Designers. New York: Dover, (1966)
8. Manjunath B. S., Ohm J. R., Vasudevan V. V. Yamada A.: Color and Texture Descriptors. IEEE Trans. Circuits and Systems for Video Technology, 11(6) 2001 703-715

Robust Video Retrieval Using Temporal MVMB Moments*

Duan-Yu Chen[1], Hong-Yuan Mark Liao[1], and Suh-Yin Lee[2]

[1] Institute of Information Science, Academia Sinica,
128 Sinica Road, Sec 2, Nankang, Taipei 11529, Taiwan
{dychen,liao}@iis.sinica.edu.tw

[2] Department of Computer Science and Information Engineering,
National Chiao-Tung University, 1001 Ta-Hsueh Rd, Hsinchu, Taiwan
sylee@csie.nctu.edu.tw

Abstract. In this paper, we propose motion pattern-based descriptor which exploits both spatial and temporal features to characterize video sequences in a semantics-based manner. The Discrete Cosine Transform (DCT) is applied to convert the high-level features from the time domain to the frequency domain. The energy concentration property of DCT allows us to use only a few DCT coefficients to precisely capture the variations of moving blobs. In comparison with the frequently used motion activity descriptors, the RLD and SAH of MPEG-7, the proposed descriptor yields 38% and 19 % average gains over RLD and SAH, respectively.

1 Introduction

In recent years, efficient retrieval of video data has become an important issue due to the popularity of the Internet. A large number of researchers have devoted themselves to the development of related core video retrieval technologies [1-6] [8-10]. Among different types of video retrieval-related techniques developed in the past decade, an automatic content analyzer that can perform efficient browsing, searching and filtering of videos is most needed. In order to design a robust content analyzer, we propose to use high-level semantic features to represent video contents. High-level video features can be derived from low-level features such as color distribution, motion distribution, or texture distribution. Among different types of low-level features that can be extracted from a video, motion is considered a very significant one due to the temporal information that it naturally owns. In the literature, Divakaran et al. [2] used a region-based histogram to compute the spatial distribution of moving regions. In [3], a run-length descriptor is proposed to reflect whether there are moving regions occurred in a frame. Aghbari et al. [4] proposed a motion-location based method to extract motion features from divided sub-fields. Peker et al. [5] calculated the average motion vectors of a P-frame and those of a video sequence to be the overall motion features. In addition to the above mentioned local motion features, Ngo et al. [6] and Tang et al. [7] proposed to use global motion features to describe video content.

In contrast to the use of motion-based features, another group of researchers proposed to use spatio-temporal features due to their abundance of data amount. In [8],

[*] This work is supported by the Ministry of Economic Affairs under contract no. 93-EC-17-A-02-S1-032

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 359–365, 2005.

Want et al. proposed to extract features from color, edge and motion, and measured the similarity between temporal patterns using dynamic programming. Lin et al. [9] characterized the temporal content variation in a shot using two descriptors - dominant color histograms of group of frames and spatial structure histograms of individual frames. Cheung and Zakhor [10] utilized the HSV color histogram to represent the key-frames of video clips and designed a video signature clustering algorithm for detecting similarities between videos. Dimitrova et al. [11] represented video segments by color super-histograms. Related works that fall into this category can be found in [12-16].

In the video comparison process, key-frame based approaches are most frequently employed [8-10] [13] [15-16]. However, there are several drawbacks associated with a matching process based on key-frames. First, the features selected from key-frames usually suffer from the high dimensionality problem. Second, the features chosen from a key-frame is in fact classified as local features. For a matching process that aims at measuring the similarity among a great number of video clips, the key-frame based matching method is not really feasible because the temporal information used to characterize the relationships among consecutive frames is not taken into account. In order to overcome the above deficiencies, we propose a motion pattern-based descriptor, which exploits the spatio-temporal information of moving blobs in the matching process. The proposed spatio-temporal features can support high-level semantic-based retrieval of videos in a very efficient manner. We make use of some spatio-temporal relationships among the moving blobs of a video clip and then use them to support the retrieval task. In the retrieval process, we use Discrete Cosine Transform (DCT) to reduce the dimensionality of the extracted high-dimensional features. Using DCT, we can maintain the local topology of a high-dimensional feature. In addition, the energy concentration property of DCT allows us to use only a few DCT coefficients to represent moving blobs and their variations.

The rest of the paper is organized as follows. Section 2 describes the methods used to characterize video segments. Section 3 presents the experiment results. Conclusions are drawn in section 4.

2 Characterizing Video Segments

In this section, we shall put our emphasis on the characterization of video segments. Since we shall build an automatic content analyzer based on motion pattern, we start from detecting the moving blobs in MPEG bitstreams. Then, the detected blobs will be represented by moments and detailed in section 2.2. In section 2.3, we shall describe how to represent the above mentioned moments across time axis.

2.1 Detecting Moving Blobs in MPEG Bitstreams

For the purpose of efficiency, the motion information associated with P-frames is used for the detection of moving blobs. In general, consecutive P-frames separated by two or three B-frames are still similar and would not vary too much. Therefore, it is reasonable to use P-frames as targets for detecting moving blobs. On the other hand, since the motion vectors embedded in MPEG bitstreams are for the purpose of compression and thus they may not be 100% correct. Under these circumstances, one has to remove the noisy part before the motion vectors can be used. In our previous work [17], a cascaded filter that is composed of a Gaussian filter followed by a median filter is exploited for noise removal. Our previous work shows that the reached precision is higher than 70% and the recall is higher than 80%. Therefore, we shall use this cascaded filter to remove noises.

2.2 MVMB Moments

After the process of noise removal, we use Motion Vector of Moving Blob (MVMB) to represent the rough features of the moving regions in a frame. Rather than directly employing the MVMB derived from a P-frame, we apply a temporal filter with window size 5 to smooth MVMBs. A basic constituent of an MVMB can be represented as *MVMB(x,y)*, which indicates the motion vector identified at location *(x,y)* in a moving blob MVMB.

To represent the spatial feature of MVMBs in a compact form, the moment invariants of MVMBs are computed. The use of moments for image analysis and object representation was propsed by Hu[19]. According to Hu's Uniqueness Theorem, the moment set { μ_{pq} } can be uniquely determined by *MVMB(x,y)* and *vice versa*. The central moment μ_{pq} computed from MVMB is defined as

$$\mu_{pq} = \sum_{y=0}^{N-1}\sum_{x=0}^{M-1}\left(x-\overline{x}\right)^p \left(y-\overline{y}\right)^q MVMB(x,y), \tag{1}$$

where *(p,q)* = {(0,2), (1,1), (2,0), (0,3), (1,2), (2,1), (3,0)} and *N*×*M* is the frame size. To select a subset of first few moments that contain sufficient information to characterize MVMBs, we adopt the top seven moment invariants defined by Hu to represent a target. These moments are as follows:

$$M_1 = \mu_{20} + \mu_{02} \tag{2}$$

$$M_2 = (\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2 \tag{3}$$

$$M_3 = (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2 \tag{4}$$

$$M_4 = (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2 \tag{5}$$

$$\begin{aligned} M_5 = (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})\left[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2\right] + \\ (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03})\left[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2\right] \end{aligned} \tag{6}$$

$$M_6 = (\mu_{20} - \mu_{02})\left[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2\right] + 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}) \tag{7}$$

$$\begin{aligned} M_7 = (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12})\left[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2\right] + \\ (\mu_{30} - 3\mu_{12})(\mu_{21} + \mu_{03})\left[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2\right] \end{aligned} \tag{8}$$

2.3 Characterizing Temporal Variations of MVMB Moments

In this section, we shall describe how to characterize the temporal variations among moving blobs based on DCT. The algorithm that can be used to generate video sequence representation is as follows:

Video Sequence Representation Algorithm

Input: Consecutive P-frames {*P1, P2, P3, …, PN*}

Output: Sequences of representative truncated DCT coefficients [$X_{\Lambda,m}$], where $\Lambda \in [1,\alpha]$, α is the number of chosen coefficients.

Procedure:

1. For each P-frame *Pi,*
   Detect moving blobs using a cascaded filter followed by using morphological operations.
2. For each P-frame *Pi,*
   Compute Hu's seven moment invariants { $M_{m,i}$ } in the filtered MVF, where $m \in [1,7]$.
3. Compute the transformed sequence [$X_{f,m}$] using the Discrete Cosine Transform
   $$X_{f,m} = C(f)\sum_{t=1}^{N} M_{m,t} cos\left(\frac{(2t+1)f\pi}{2N}\right), \text{ where } f \in [1,N]$$
4. For *m* transformed sequences, [$X_{f,m}$] of DCT coefficients,
   Truncate the number of DCT coefficients to α, which is composed of the DC coefficient and (α-*1*) AC coefficients to represent a transformed sequence.
5. 5.Generate a feature vector *F(* $X_{\Lambda,1}, X_{\Lambda,2}, X_{\Lambda,3}, X_{\Lambda,4}, X_{\Lambda,5}, X_{\Lambda,6}, X_{\Lambda,7}$ *)* for each video segment, where $\Lambda \in [1,\alpha]$.

In the above algorithm, the spatial feature of moving blobs in a P-frame is represented by Hu's seven moment invariants. In order to characterize the temporal variations of moving blobs among successive frames, DCT is exploited to transform the MVMB moments of the original video sequence into the frequency domain. We use $M_{m,i}$ to represent the value of MVMB at the *i*th P-frame. This value can be considered as a signal appeared at time *i*. For N consecutive P-frames, their MVMB values can be expressed as a sequence of signals x_m = [$M_{m,t}$], where *t = 1, 2,…, N*. The N-point DCT of a signal x_m is defined as a sequence *X* = [$X_{f,m}$], *f= 1, 2, 3, …, N* as follows:

$$X_{f,m} = C(f)\sum_{t=1}^{N} M_{m,t} cos\left(\frac{(2t+1)f\pi}{2N}\right), \tag{9}$$

$$C(0) = \sqrt{\frac{1}{N}} \text{ and } C(f) = \sqrt{\frac{2}{N}},\ f = 1,2,..,N-1$$

where *N* is the number of P-frames and *m* ∈ [1,7]. Eq. (9) indicates that a video sequence is represented by *7* sequences of DCT coefficients. It means that the temporal variations existing among the original objects in the set of successive P-frames are characterized by *7* sequences of DCT coefficients in the frequency domain. It is well known that the first few low-frequency AC terms together with the DC term will have enough representation power. Therefore, for considering computation cost we only choose these terms to represent a video sequence instead of selecting all coefficients. However, to select an appropriate amount of AC coefficients is always a crucial issue. The statistics computed from a large number of experiments has indicated that two DCT coefficients are good enough for comparing the similarity among different video segments.

3 Experiment Results

3.1 Similarity Measure

The similarity measure is for computing the similarity between a query video shot and a model video shot. In order to choose an appropriate similarity measure, we took the variability of each component that would be compared into account. For those components that have high variability should be associated with less weight than those components who have low variability. Under these circumstances, the Mahalanobis distance is used as a similarity measure, which is defined as

$$D\left(F^q, F^t\right) = \left(\sum_{k=1}^{n} \left| \frac{F_k^q - F_k^t}{\sigma_k} \right|^2 \right)^{1/2}, \tag{10}$$

where F_k^q and F_k^t denote the kth component of a query feature vector F^q and a model feature vector F^t, respectively. n here denotes the dimension of a feature vector and σ_k denotes the standard deviation of the kth component of all feature vectors in the test dataset.

3.2 Performance Evaluation

In order to demonstrate the effectiveness of the proposed method, we simulated the algorithm of video sequence matching by using MPEG-7 test dataset [18] which included various kinds of videos, such as news, sports, entertainment, education, etc. The total number of shots included in the test set was 1027. The degree of motion strength of these videos ranged from low, medium to high, and the size of every moving object in the video was classified as either small, medium or large. To evaluate the performance, precision and recall were used as the metrics to measure the performance of the proposed system.

In the experiments, we used three classes of shots to test the performance of our algorithms. Among these test videos, the shots covered in the Close-Up Tracking (CUT) video and the Walking Persons (WP) video were with high degree of motion and medium degree of motion, respectively. The shots belonging to the Interview (IV) video were with low degree of motion. The sensitivity of the proposed descriptor is dependent on the size of moving blobs. For those blobs that were smaller than 2x2 macroblocks, we considered them small blobs. For those blobs whose area occupied more than half of a frame, we considered them large blobs. For each query process, we retrieved the top 30 shots from the model database. In order for making to comparison, we conducted the same set of experiments using two other algorithms. They were motion-based run-length descriptor (RLD) and shot activity histogram (SAH) provided by MPEG-7. Fig. 1 shows the curves of the precision versus the recall calculated by three different algorithms. Our algorithm obtained 45% average gain in the IV video, 30% in the CUT video and 34% in the WP video in comparison with the RLD algorithm. In average, the proposed descriptor yielded 38% and 19% average gain in comparison with the RLD algorithm and the SAH algorithm, respectively.

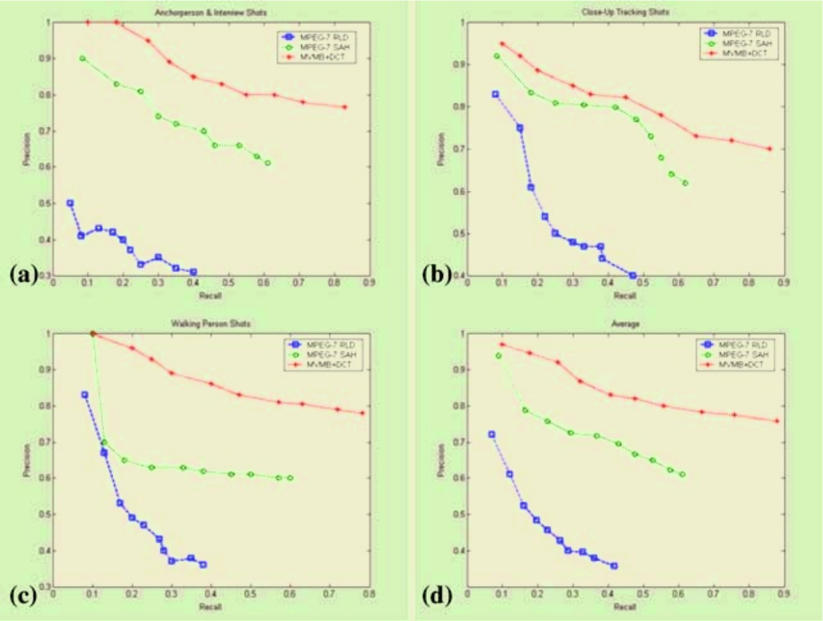


Fig. 1. The curves showing the recall versus the precision calculated by the three different algorithms (a) Interview Shots (b) Close-Up Tracking Shots (c) Walking Person Shots (d) Average

4 Conclusions

A novel motion pattern-based descriptor for video sequence matching has been developed in this work. The proposed descriptor has two special features: 1) the proposed temporal MVMB moments has exploited both spatial and temporal features of moving blobs and characterized video sequences in a high-level manner; 2) the dimensionality of feature space has been reduced using DCT while characterizing the temporal variations among moving blobs. Experiment results have demonstrated that a few DCT coefficients could suffice for representing a video sequence and also shown that the proposed motion-pattern descriptor was quite robust. Using this descriptor, one can perform video retrieval in an accurate and efficient way.

References

1. T. Sikora, “The MPEG-7 Visual Standard for Content Description – An Overview,” IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No. 6, pp. 696 –702, June 2001.

2. ADivakaran, K. Peker and H. Sun, "A Region Based Descriptor for Spatial Distribution of Motion Activity for Compressed Video," Proc. International Conference on Image Processing, Vol. 2, pp. 287-290, Sep. 2000.
3. S. Jeannin and A. Divakaran, "MPEG-7 Visual Motion Descriptors," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No. 6, pp. 720-724, June 2001.
4. Z. Aghbari, K. Kaneko and A. Makinouchi, "A Motion-Location Based Indexing Method for Retrieving MPEG Videos," Proc. 9th International Workshop on Database and Expert Systems Applications, pp. 102-107, Aug. 1998.
5. K. A. Peker, A. A. Alatan and A. N. Akansu, "Low-Level Motion Activity Features for Semantic Characterization of Video," Proc. IEEE International Conference on Image Processing, Vol. 2, pp 801-804, Sep. 2000.
6. C. W. Ngo, T. C. Pong and H. J. Zhang, "On Clustering and Retrieval of Video Shots," Proc. ACM Multimedia Conference, pp. 51-60, Ottawa, Canada, Oct. 2001.
7. Y. P. Tang, D. D. Saur, S. R. Kulkarni and P. J. Ramadge, "Rapid Estimation of Camera Motion from Compressed Video with Application to Video Annotation," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 10, No. 1, pp. 133-146, Feb. 2000.
8. R. Wang, M. R. Naphade, and T. S. Huang, "Video Retrieval and Relevance Feedback in The Context of A Post-Integration Model," Proc. IEEE 4th Workshop on Multimedia Signal Processing, pp. 33-38, Oct. 2001.
9. T. Lin, C. W. Ngo, H. J. Zhang and Q. Y. Shi, "Integrating Color and Spatial Features for Content-Based Video Retrieval," Proc. IEEE International Conference on Image Processing, Vol. 2, pp. 592-595, Oct. 2001.
10. S. S. Cheung and A. Zakhor, "Video Similarity Detection with Video Signature Clustering," Proc. IEEE International Conference on Image Processing, Vol. 2, pp. 649–652, Sep. 2001.
11. L. Agnihotri and N. Dimitrova, "Video Clustering Using SuperHistograms in Large Archives," Proc. 4th International Conference on Visual Information Systems, pp. 62-73, Lyon, France, November 2000.
12. M. Roach, J. S. Mason and M. Pawlewski, "Motion-Based Classification of Cartoons," Proc. International Symposium on Intelligent Multimedia, Video and Speech Processing, pp. 146-149, Hong Kong, May 2001.
13. L. Zhao, W. Qi, S. Z. Li, S. Q. Yang and H. J. Zhang, "Content-based Retrieval of Video Shot Using the Improved Nearest Feature Line Method," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 3, pp. 1625-1628, 2001.
14. B. S. Manjunath, J. R. Ohm, V. V. Vasudevan and A. Yamada, "Color and Texture Descriptors," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No. 6, pp. 703-715, June 2001.
15. R. Mohan, "Video Sequence Matching," IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 6, pp. 3697-3700, May 1998.
16. M. M. Yeung and B. Liu, "Efficient Matching and Clustering of Video Shots," Proc. IEEE International Conference on Image Processing, Vol. 1, pp. 338-341, Oct. 1995.
17. Ahmad, A.M.A., D. Y. Chen and Suh-Yin Lee, "Robust Object Detection Using Cascade Filter in MPEG Videos," Proc. IEEE 5th International Symposium on Multimedia Software Engineering, pp. 196-203, Taichung, Taiwan, Dec 2003.
18. ISO/IEC JTC1/SC29/WG11/N2466, "Licensing Agreement for the MPEG-7 Content Set," Atlantic City, USA, October 1998.
19. M. Hu, "Visual Pattern Recognition by Moment Invariants," IRE Transactions on Information Theory, Vol. IT-8, pp. 179-187, Feb. 1962.

Precise Segmentation Rendering for Medical Images Based on Maximum Entropy Processing

Tsair-Fwu Lee[1,2], Ming-Yuan Cho[1], Chin-Shiuh Shieh[4],
Pei-Ju Chao[3], and Huai-Yang Chang[2]

[1] Department of Electrical Engineering, National Kaohsiung University of Applied Science, Kaohsiung, Taiwan 807, ROC

[2] Department of Radiation Oncology, Chang Gung Memorial Hospital-Kaohsiung, 83305, Taiwan, ROC

[3] Department of Radiation Oncology, Kaohsiung Yuan's General Hospital, Kaohsiung, 800, Taiwan, ROC

[4] Department of Electronic Engineering, National Kaohsiung University of Applied Science, Kaohsiung, Taiwan 807, ROC

Abstract. Precision is definitely required in medical treatments, however, most three-dimensional (3-D) renderings of medical images lack for required precision. This study aimed at the development of a precise 3-D image processing method to discriminate clearly the edges. Since conventional Computed Tomography (CT), Positron Emission Tomography (PET), or Magnetic Resonance Imaging (MRI) medical images are all slice-based stacked 3-D images, one effective way to obtain precision 3-D rendering is to process the sliced data with high precision first then to stack them together carefully to reconstruct a desired 3-D image. A recent two-dimensional (2-D) image processing method known as the entropy maximization procedure proposed to combine both the gradient and the region segmentation approaches to achieve a much better result than either alone seemed to be our best choice to extend it into 3-D processing. Three examples of CT scan data of medical images were used to test the validity of our method. We found our 3-D renderings not only achieved the precision we sought but also has many interesting characteristics that shall be of significant influence to the medical practice.

Keywords: segmentation, wavelet, edge detection, entropy maximization.

1 Introduction

Physicians need 3-D renderings to help them to make diagnosis, conduct surgery, and provide other treatments that 2-D images and other conventional test methods cannot offer. Without precise segmentation, renderings could lead to misleading results. The aim of this study is to provide a precise 3-D rendering method to achieve what physicians demand. Two basic approaches in existing works on image segmentation are: the gradient-based approach and the region-based approach. Gradient-based edge detection methods [1,2] rely on the local

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 366–373, 2005.

differences in the gray scale values of an image. They focus on the differences and transitions of the intensity of an image. The disadvantage in these methods is that they almost always result in broken and false edges. Region-based segmentation techniques include region-growing, region-splitting, and region-merging algorithms, *etc.* focused on the homogeneity of spatially dense localized features and other pixel statistics. They have a common problem of over-segmented, and hence produces poorly localized region boundaries. To resolve their weaknesses and to combine their strengths of the gradient-based approach and the region-based approach, Staib and Duncan proposed an idea to combine both of them in the maximum entropy manner to achieve a better result [3,4]. And recently Duncan *et al.* [5] have some successful applications in its extented works. Moreover, some authors paid more attention in this field with different methods in progress [6].

In this study, our objective is to apply the Staib and Duncan method [3,4] to combine various segmentation approaches with an entropy maximization procedure and extended the idea into 3-D area. This allows us to utilize all available information to achieve the most robust segmentation results for 3-D image processing. We then apply our combined segmentation method to medical images to test the validity of our method. We aim to show our combined 3-D segmentation method is indeed superior in terms of required precision, also the sliced-base approach we proposed is quite efficient, and many features generated by our 3-D segmentation method shall be of referential values to the physicians.

2 Wavelet Edge Detector

No doubt the Wavelet method is known to be one of the best gradient segmentation methods due to its multi-scale and multi-resolution capabilities. We briefly discuss some of its property in this section. We shall name $S_{2^j}[.]$ and $D_{2^j}[.]$ as the low pass signal (or the approximated signal) and the high pass signal (or the detailed signal) of $f(x)$ at resolution 2^j**,** respectively. And $S_{2^j}[n]$ is the projection coefficient of $f(x)$ on subspace V_j, $D_{2^j}[n]$ is the projection coefficient of $f(x)$ on subspace O_j.We can define an orthogonal complement subspace of V_j as O_j, in space V_{j+1}. The scaling function $\phi(x)$ and wavelet function $\varphi(x)$ have the orthogonal properties. From the properties of multiresolution analysis, we can easily see that signals can always be decomposed into higher resolutions until the desired result is obtained. A 2-D filter for edge detection is generated by a 2-D discrete periodic wavelet transform (2-D DPWT) [4], applying separable algorithms, the 2-D DPWT can be written in a matrix form. And we now extend the wavelet transform into two dimensional manners. So we can define the four operators [4,7] of 2-D DPWT as follows: (Reader can refer to the detail description in the reference [8] which proposed by Mallat in 1989.)

$$W_{LL} = [h(i) \cdot h(j)]_{i,j \in Z} \tag{1}$$

$$W_{LH} = [(-1)^{3-j} h(i) \cdot h(3-j)]_{i,j \in Z} \tag{2}$$

$$W_{HL} = [(-1)^{3-j} h(3-i) \cdot h(j)]_{i,j \in Z} \tag{3}$$

$$W_{HH} = [(-1)^{i+j} h(3-i) \cdot h(3-j)]_{i,j \in Z} \tag{4}$$

where, W_{LL}, W_{LH}, W_{HL} and W_{HH} are the four subband filters; and the $\otimes$ denoted a convolution operation; and $h(i) =< \phi_{2^{-1}}(u) \cdot \phi(u-i) >$. Clearly, as the length of the coefficients of the filter is d, the operator of 2-D DPWT formed a $d \times d$ matrix. We now use the coefficients of the four filters given by Eq. 1 to Eq. 4 to generate a wavelet edge detector. Let $f_h(i,j)$ be the horizontal high-pass filter function and $f_v(i,j)$ be the vertical high-pass filter function. These two high-pass filters are obtained from the four operators of 2-D DPWT

$$f_h(i,j) = W_{LL}(i,j) \otimes W_{LH}(i,j) \tag{5}$$

$$f_v(i,j) = W_{LL}(i,j) \otimes W_{HL}(i,j). \tag{6}$$

We then apply the different length coefficients of Daubechies wavelet transform to generate the multi-scale masks (filters)[4,7,8].

Therefore, let the original image pass through these masks to produce a series of multi-scale images with different gradient strength scales. In order to avoid distortions caused by noise and to define exact edge points, an edge thinning technique is then used to make effective combinations of the images [9].

3 Concepts of Maximum Entropy Processing

By boundary estimation we meant to find optimum values of the boundary parameters as the information of image data were given. Let us define the optimization of the entropy function by maximizing its *a posteriori* probability (MAP) [3,10]. Assuming that $I_b(x,y)$ is the image that depicts some object and $t_{\hat{p}}(x,y)$ is the image template corresponding to the parameter vector $\hat{p}$. We maximize $P(t_{\hat{p}}|I_b)$, the conditional probability of the template given the image, to obtain the best estimate of $\hat{p}$. By Baye's rule, the function $P(t_{\hat{p}}|I_b)$ can be written as follows:

$$\arg\max_{\hat{p}} P(t_{\hat{p}}|I_b) = \arg\max_{\hat{p}} \frac{P(I_b|t_{\hat{p}})P(t_{\hat{p}})}{P(I_b)} \tag{7}$$

where $P(t_{\hat{p}})$ is the *a priori* probability of the template and $P(I_b|t_{\hat{p}})$ is the conditional probability of image I_b, which depicts some object with template $\hat{p}$. The denominator of Eq. 7 is not a function of $\hat{p}$ and hence can be ignored. Taking logarithm of Eq. 7 we have:

$$\arg\max_{\hat{p}} M(I_b, t_{\hat{p}}) = \arg\max_{\hat{p}} [lnP(t_{\hat{p}}) + lnP(I_b|t_{\hat{p}})] \tag{8}$$

To estimate the parameter vector $\hat{p}$ we maximize the entropy function $M(I_b, t_{\hat{p}})$. Clearly, the first term of Eq. 8 represents the *a priori* information and the second term represents the data-driven likelihood term. After rearranged the equation by Baye's rule, we find the boundary estimation problem becomes

$$\arg\max_{\hat{p}} ln[P(\hat{p}|I_g, I_r)] \equiv \arg\max_{\hat{p}} [lnP(\hat{p}) + lnP(I_g|\hat{p}) + lnP(I_r|(I_g, \hat{p})] \tag{9}$$

Clearly, the first term is the *a priori* information, the second term contains the gradient-based information I_g, and the last term is the region-based informationI_r

conditioned on parameter vector $\hat{p}$ and the gradient-based information I_g. We argue that we shall ignore the dependence on I_g since the information of gradient-based information is already obtained by the second term. Then Eq. 9 becomes

$$\arg\max_{\hat{p}} M(\hat{p}, I_g, I_r) \equiv \arg\max_{\hat{p}} [M_{prior}(\hat{p}) + M_{gradient}(I_g, \hat{p}) + M_{region}(I_r, \hat{p})] \tag{10}$$

For each slice, the input consists of the original image I and the region-classified image I_r which is the result of region-based segmentation as discussed above. Next a gradient-based approach is added and it uses the gradient image I_g. As described by Staib and Duncan [3,10], we shall use the magnitude of the gradient vector at each voxel location. I_g can be obtained from I either by convolving the input image I with the multiple wavelet edge detection operators and then computing I_g to be the magnitude of the above resulting vector image. Hence the input to the system is the gradient image I_g and the region-classified image I_r. Only when slice processing is completed, we use linear interpolation algorithm along the slices direction to form 3-D surfaces. We argue that, our surface estimation method using both gradient and region homogeneity information is still in the maximum *a posteriori* framework. We have suitably incorporated *a priori* shape information when region-of-interest (ROI) is available.

4 Medical Images Applications and Results

The goal of our precise image segmentation is to partition an image into disjoint regions of desired objects as accurate as possible. Among the proposed segmentation methods, region-growing has been the most popular one due to its speed and great flexibility. We use symmetric seeds [11] to initiate the segmentation to avoid a single seed may fall into a noise region too easily. Then we combine the region-growing segmentation with the gradient-based segmentations, namely the wavelet edge detector. The following are precision 3-D renderings generated by our combine-information segmentation applies to medical images of distinct characteristics and medical importance.

4.1 Medical Image Experiment 1

A set of CT scan, slices of a human chest which we want to process only a particular region of medical interests. This also demonstrated the great capability and flexibility of our combine-information segmentation method.

We find the edges by following the boundary-finding procedure. The gradient-based information is obtained first by using the wavelet edge detection method, and the region-based information is obtained from the region-growing method initialized with symmetric seeds. The entropy function Eq. 10 completes the combination in maximum sense. As we have discussed earlier, when we process the gradient-based information only, the first two terms are used as entropy functions. As we process the image with the region-based information only, we used the first and the third terms as entropy functions. Finally, all three terms are used as entropy functions to complete the combination. We shall perform the segmentation first slice by slice. Since we aim to extract the right lung for feature

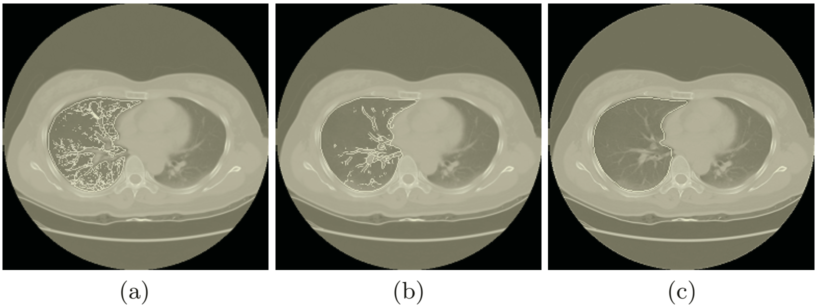


Fig. 1. (a) Segmented contour by the region-growing method. (b) Segmented contour by the wavelet method. (c) Segmented contour by the combination method

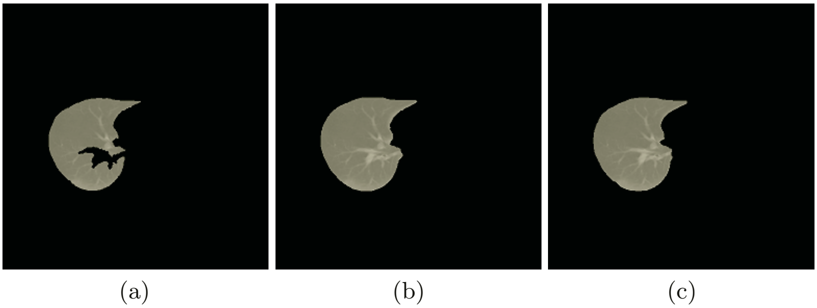


Fig. 2. (a) Segmented result of the region-growing method.(b) Segmented result of the wavelet method. (c) Segmented result of the combination method

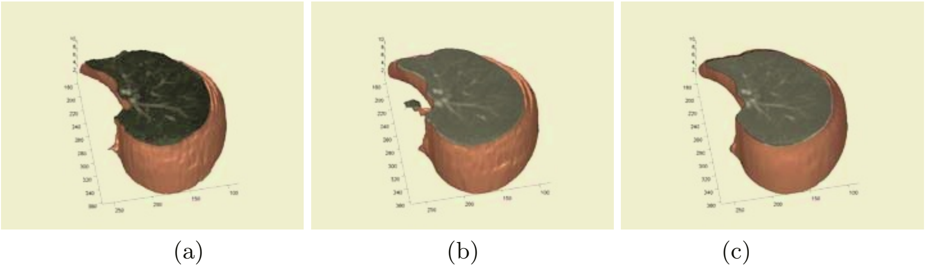


Fig. 3. (a) 3-D rendering by the region-growing method. (b) 3-D rendering by wavelet method. (c) 3-D rendering by the combination method

extraction in this case hence it becomes our natural ROI. Fig. 1 (a) shown the region-growing contour, Fig. 1 (b) is the wavelet contour, and Fig. 1 (c) is the combined contour. Figs. 2 (a) (b) and (c) present the segmentation results of Figs. 1 (a) (b) and (c) respectively. On close inspection of these figures, shortcomings from either segmentation although were obvious but seemed harmless. However, as 3-D renderings were form, they become serious errors, and shall not be tolerated if precision renderings were sought.

Fig. 3 (a) shows the 3-D rendering from the region-growing segmentation. Fig. 3 (b) shows the 3-D rendering from the wavelet segmentation. Fig. 3 (c) shows the 3-D rendering from the combined segmentation. If we look at them closely, we found that the 3-D renderings either by the region-growing segmentation, or by the wavelet segmentations, although seemed both acceptable, but they both have problems that at times a single slice will be very different from the others at some particular point, perhaps due to noise and/or other disturbances, which makes the corresponding 3-D renderings appeared with wrinkles. Clearly human lungs should be continuous and smooth in all directions always, hence we may conclude that both the region-growing method and the wavelet method fail to reconstruct the object precisely.

As expected, the maximum entropy combination results in a 3-D rendering with much better quality. Hence we can conclude that the combined method is indeed superior to each individual processing alone, and our purpose of seeking precision has been achieved.

4.2 Medical Image Experiment 2

Next we shall test the accuracy of localization of our method. By the use of coloring and a so-called transparency technique, not only we can identify the problem area precisely, but also its relative position to other critical organs. The test data were the CT scan of a female patient with a pituitary tumor in her brain. The pituitary gland is about the size of a pea in the center of our brain just at the back of our nose. The choice of treatment uniquely depends on the position of the tumor.

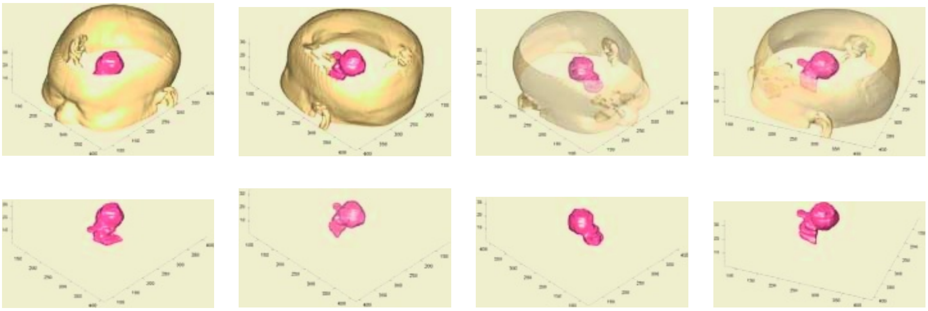


Fig. 4. 3-D renderings of the tumor and the head in two colors, four angles, and transparent effect

Here we emphasize the importance of precision. With a target so small and so vital, only position of the tumors can be pin down with the highest precision, treatments can then be effective, and ordinary brain cells shall not be damage.

In order to identify the tumor clearly, we first segmented it out with great precision and then 3-D renderings of the pituitary tumor are then constructed. The tumor can now be clearly inspected by checking the Fig.4. The figures clearly demonstrated the power of our segmentation, capable of providing an outstanding positioning of the tumor, which has not been achieved by other 3-D renderings previously. The size and shape of the tumor, its orientation with the brain, and the position relative to the head are all now clearly seen.Its various angles are shown in the Figs. These 3-D renderings can be rotated to any angle and with different colors for physicians to inspect closely. Transparency effect is now introduced which shall be most useful for radiation therapies.

4.3 Medical Image Experiment 3

Finally we shall present the 3-D renderings for a common orthopedic decease happens to both genders for senior citizens. Human stands on two feet and the joint between our legs and our pelvis solely supports our weight. In medical terms, it is the femur connecting to the acetabulum supporting our weight. Over years of overly use, the joint becomes rough and causes pain. If it is not taken care of properly, fragment shall occur. How were the joint over used, what are the damages, in present days orthopedic surgeon either depends on CT scan or inspect by endoscopes. However, endoscopes inspection is time consuming; CT slices although were the best X-ray scan able to provide, really did not fully expose the problem.

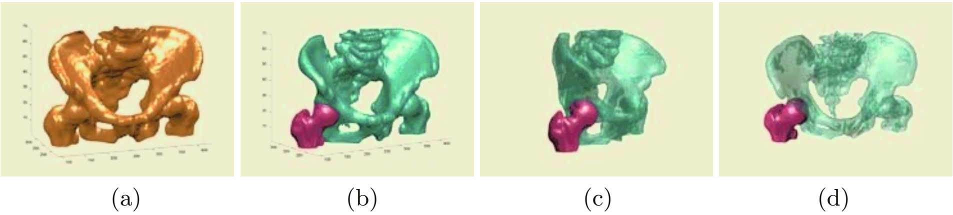

(a) (b) (c) (d)

Fig. 5. (a) 3-D rendering of the pelvis. (b) (c) 3-D renderings of the right joint in different angle. (d) The pelvis with transparent effect in different angle

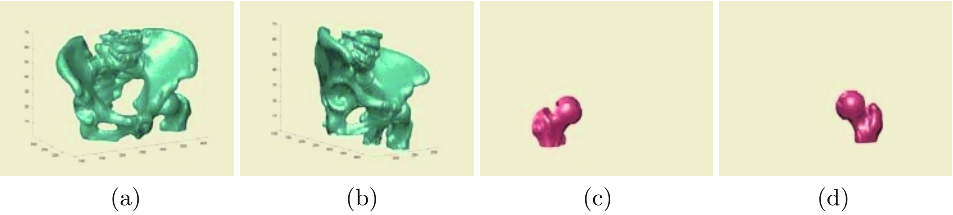

(a) (b) (c) (d)

Fig. 6. (a) 3-D rendering of the acetabulum in pelvis. (b) The acetabulum in pelvis in a different angle. (c) (d) 3-D renderings of the femur in different angles

With the reconstruction techniques we have developed, precision 3-D renderings of all angles, enlargements, distinct part of the joint, can all be visualize much clearly allowing physicians to make correct decisions. Fig. 5 (a) shows the pelvis only for physicians to inspect in detail. Fig. 5 (b) (c) shows the complete right joint of the pelvis in different angles. Figs. 5 (d) showed its different angle with transparent effect to simulate various movements of the joint. Figs. 6 (a) and (b) are the close look of the acetabulum. Fig. 6 (c) (d) shows the different angle of the femur. With the help of these 3-D renderings, physicians shall be able to make diagnosis more efficiently and effectively.

5 Conclusion

On all examples of medical images we processed, not only desired precision had been achieved, we are also able to create rotation of the objects to obtain its 3-D images of different angles. The 3-D renderings we created will allow physicians to conduct surgery or treatment much more accurately and effectively. Many images of interest that physicians unable to visualize, but have to compose a 3-D image by their imaginations, all become possible after our 3-D processing. Features are now clearly identified, locations pinned down exactly, and relative orientations are now well understood. These are all vital for medical treatments.

Therefore we may conclude that our 3-D rendering method that combines the gradient-based and the region-based information in the maximum entropy sense, not only proved to be a superb image processing techniques but also very useful in practice for medical images. We believe that our precision 3-D renderings shall play its role in future medical applications.

References

1. John C. Russ. The Image Processing Handbook, Third ed. CRC Press & IEEE Press, 1999.
2. Rafael C. Gonzalez, Richard E. Woods. Digital Image Processing. Prentice Hall, 2^{nd} ed. Edition, 2002.
3. L.H. Staib. “Boundary finding with parametrically deformable models.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.14, no.11, pp.1061-1075, 1992.
4. Cheng-Tung Ku, King-Chu Hung and Mig-Cheg Liag. “Wavelet Operators for Multi-scale Edge and Corner Detection.” Department of Electrical Engineering, I-Shou University, Taiwan, 1998.
5. Jing Yang, James S. Duncan. “Joint Prior Models of Neighboring Objects for 3D Image Segmentation”, Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’04), 1063-6919/04,2004.
6. Hua Li, Abderr Elmoataz, Jalal Fadili, Su Ruan, Barbara Romaniuk. “3D Medical Image Segmentation Approach Based on Multi-Label Front Propagation “, IEEE 2004 International Conference on Image Processing (ICIP), pp. 2925-2928, 2004.
7. Yih-Sheng Leu, Chao-Ji Chou. “Wavelet Edge Detection on Region-based Image Segmentation” Department of Computer & Communication Engineering, National Kaohsiung First University of Science and Technology, Taiwan, 2000.
8. S.G. Mallat, “A Theory for Multiresolution Signal Decomposition: The Wavelet Representation.” *IEEE Transactions on Analysis and Machine Intelligence*, vol. 11. no. 7, 1989.
9. Gabriele Lohmann. Volumetric Image Analysis. Wiley & Teubner, 1998.
10. A. Chakraborty, “Feature and Module Integration for Image Segmentation.” Ph.D thesis, Yale University, 1996.
11. Shu-Yen Wan, William E. Higgins. “Symmetric Region Growing.“ *IEEE Transactions on image processing*, vol. 12, no.9, pp 1007-1015, September 2003.

A Hardware Implementation for Fingerprint Retrieval

Yongwha Chung[1], Kichul Kim[2], Min Kim[2], Sungbum Pan[3], and Neungsoo Park[4]

[1] Korea University, Korea
ychungy@korea.ac.kr
[2] University of Seoul, Korea
{kkim,minkim}@uos.ac.kr
[3] Chosun University, Korea
sbpan@chosun.ac.kr
[4] Konkuk University, Korea
neungsoo@konkuk.ac.kr

Abstract. Recently, fingerprint has been received considerable attentions as a user authentication technique using biometrics. Fingerprint retrieval, which retrieves a candidate list of fingerprints having similar features with the given fingerprint from a fingerprint database, is one of interesting real-time applications. However, using the straightforward approach, it takes a long response time to scan the entire database and to compare the query against each reference on a large-scale database. Even when implemented on a hardware, it is hard to satisfy the real-time requirement. In this paper, to reduce the response time, we employ the geometric hashing technique which has been widely used for model-based object recognition. Using this algorithm, the number of fingerprint minutiae can be restricted. It limits the hardware size used for implementation to fit into FPGA-chips. Furthermore, we develop a scalable implementation for parallel geometric hashing on an FPGA-based PCI platform.

Keywords: Fingerprint Retrieval, Real-Time Processing, Special-Purpose Hardware

1 Introduction

Traditionally, verified users have gained access to secure information systems, buildings, or equipment via multiple PINs, passwords, smart cards, and so on. However, these security methods have critical weakness that can be lost, stolen, or forgotten. Recently, **biometrics** has been received considerable attentions, which refers the personal biological or behavioral characteristics used for verification or identification.

In this paper, the **fingerprint** has been chosen as the biometrics for verification. It is more mature in terms of algorithm availability and feasibility[1]. We focus on the fingerprint **retrieval** or **identification**(*one-to-many* matching) on large-scale databases rather than the fingerprint verification(*one-to-one* matching). The retrieval problem is more challenging because it necessitates a large database search of individuals to determine whether a person is already in the database.

The traditional fingerprint retrieval system has been used for **criminal** justice applications. However, this paper targets for **non-criminal** applications. Examples include a social service database, wherein individuals must be prevented from using multiple aliases, watch list check in an immigration office, and identity card issuance.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 374–380, 2005.

The main difference of those two applications is the response time. Retrievals in non-criminal applications need to be completed in few seconds, whereas few hours/days are acceptable in criminal applications.

The straightforward approach to search a large database is to scan the entire database and to compare the query against each reference model. Even though this approach could be implemented by hardware, the real-time requirement can not be satisfied for large-scale databases[2]. To speed-up the response time, we employ the **geometric hashing** technique[3] which has been used for model-based object recognition. By generating a fingerprint index table carefully in the enrollment phase, we can perform the retrieval without time-consuming alignment process.

To speed-up further, we develop a **parallel** geometric hashing on an **FPGA-based PCI platform**. There have been some prior efforts in parallelizing the geometric hashing algorithm on parallel machines[4]. A major problem in these implementations is that their performance degrades due to the irregular pattern in accessing the hash bins and in voting. Also, some FPGA-based geometric hashing solutions have been proposed for target recognition[5]. Although they can work well for fine-grain applications such as target recognition, they have some limitation in large-scale applications such as fingerprint retrieval. In this paper, we propose a scalable solution which provides both fine-grain(within a board) and coarse-grain(between boards) level parallelisms for real-time fingerprint retrieval. The proposed design was implemented on PCI boards with FPGAs and SDRAMs. Based on experimental results, we confirm that the real-time performance for large-scale fingerprint retrieval can be provided using multiple PCI boards proposed in this paper.

The rest of the paper is structured as follows. Section 2 explains the overview of a typical fingerprint retrieval process and the geometric hashing, and Section 3 describes the proposed fingerprint retrieval based on geometric hashing. The experimental results are given in Section 4, and conclusions are made in Section 5.

2 Fingerprint Retrieval

This section explains enrollment process first in detail. In enrollment process, minutiae information of a user is extracted and a table, called an *enrollment table*, is generated according to the geometric characteristics of the minutiae. Enrollment process consists of minutiae information acquisition stage, table generation stage, and fingerprint database update stage. In minutiae information acquisition stage, minutiae are extracted from the fingerprint image of an enrollment user. A minutia can be specified by its coordinates, angle, and its type. Let $M_i = (x_i, y_i, \theta_i, t_i)$ represent a minutia. The coordinates show the position of the minutia. The angle shows the direction of the minutia. Finally, the type shows if the minutia is an ending point or a bifurcation point. An enrollment user can be represented by the set of minutiae and can be represented by $E = \{M_i \mid 0 \le i \le n - 1\}$.

The geometric characteristics of minutiae of a user varies over fingerprint images. A fingerprint image can be translated, rotated, enlarged or shrinked in each extraction. Hence, a direct comparison between two fingerprint images is impossible. Alignment is an essential step when comparing a fingerprint with fingerprints in the fingerprint database. Alignment is a very time consuming process. In this paper, enrollment tables are generated in such a way that no alignment is needed in the identification process by using the geometric hashing technique[3]. In other words, alignment is pre-performed in the enrollment table generation stage and the results are stored in the fingerprint database. In identification process, direct comparisons without alignment are performed in 1:*N* matching between an identification user fingerprint with fingerprints in the database.

Table generation stage is a stage in which an enrollment table is generated from input minutiae. Each step in the table generation stage is explained in detail in the following.

1) Reference Point Selection Step
In reference point selection step, a minutia is selected as the first minutia from the set of enrolment user minutiae. The first minutia is denoted by M_0 and the other remaining minutiae are denoted as M_1, M_2, …, M_{n-1}.

2) Minutiae Transform Step
In minutiae transform step, minutiae M_1, M_2, …, M_{n-1} are aligned with respect to the first minutia M_0 and quantized. Let M_j^0 denote the transformed minutiae, *i.e.*, the result of the transform of the *j*th minutia with respect to M_0. Also, let T_0 be the set of transformed minutiae M_j^0, *i.e.*, $T_0 = \{M_j^0 = (x_j^0, j_j^0, \theta_j^0, t_j^0) \mid 0 < j \le n-1\}$, and T_0 is called the M_0-transformed minutiae Set. Eq. 1 performs translation and rotation such that features $(x_0, y_0, \theta_0, t_0)$ of M_0 is translated and rotated into $(0,0,0,t_0)$. ${}_{TR}M_j^0$ is the minutia translated and rotated from the *j*th minutia with respect to M_0.

$$ {}_{TR}M_j^0 = \begin{pmatrix} {}_{TR}x_j^0 \\ {}_{TR}y_j^0 \\ {}_{TR}\theta_j^0 \\ {}_{TR}t_j^0 \end{pmatrix} = \begin{pmatrix} \cos(\theta_0) & \sin(\theta_0) & 0 & 0 \\ -\sin(\theta_0) & \cos(\theta_0) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_j - x_0 \\ y_j - y_0 \\ \theta_j - \theta_0 \\ t_j \end{pmatrix}, where\ 0 < j \le n-1 \tag{1} $$

$$ M_j^0 = \begin{pmatrix} x_j^0 \\ y_j^0 \\ \theta_j^0 \\ t_j^0 \end{pmatrix} = \begin{pmatrix} \lfloor {}_{TR}x_j^0 / \alpha + 0.5 \rfloor \\ \lfloor {}_{TR}y_j^0 / \alpha + 0.5 \rfloor \\ \lfloor {}_{TR}\theta_j^0 / \beta \rfloor \\ {}_{TR}t_j^0 \end{pmatrix} \tag{2} $$

To reduce the amount of information, quantization is performed both on coordinates and angles as shown in Eq. 2. α is the quantization parameter for coordinates and β is the quantization parameter for angles. The quantization parameter for coordinates is determined by the range of coordinates in the extraction stage. The quantization parameter for angles is determined by the required precision in the identification process.

3) Repeat Step
Step 1) and step 2) are repeated for all remaining minutiae. When step 1) and step 2) are finished for all minutiae of the enrollment user, enrollment table is completed. In fingerprint database update stage, enrollment table is added to the fingerprint database.

After enrollment process, identification process should be performed. In identification process, minutiae information of an identification user is obtained and a table, called *identification table*, is generated according to the geometric characteristic of the minutiae. Then, the identification table is compared with the fingerprint database, and finally a candidate list is generated according to the level of similarity.

Identification process consists of minutiae information acquisition stage, table generation stage, 1:*N* matching stage, and candidate list generation stage. Of these stages, minutiae information acquisition stage and table generation stage are performed in the same way as in the enrollment process. An identification user can be represented by the set of minutiae. Let $N_i, 0 \le i \le n-1$, represent the minutiae of the identification user, and let I represent the set of minutiae of the identification user, *i.e.*, $I = \{N_i \mid 0 \le i \le n-1\}$. An identification table is generated in the same way an enrollment table is generated.

In 1:*N* matching stage, the number of transformed minutiae pairs with the same coordinates, the same angle, and the same type are computed. For the purpose of explanation, let $N_1^0, N_2^0, N_3^0, N_4^0, N_5^0, N_6^0$ be N_0-transformed minutiae in identification user fingerprint and $M_1^0, M_2^0, M_3^0, M_4^0, M_5^0, M_6^0$ be M_0-transformed minutiae in the fingerprint database. Let's assume all minutiae are of the same type. In this example, coordinates and angles match exactly in three pairs of minutiae, *i.e.*, (N_1^0, M_1^0), (N_2^0, M_2^0), (N_3^0, M_3^0). There is no exact match of coordinates and angles in minutiae $N_4^0, N_5^0, N_6^0, M_4^0, M_5^0, M_6^0$. As shown in this example, similarity between two fingerprints can be measured by counting the number of matched transformed minutiae pairs. Using transformed minutiae table and transformed minutiae database, 1:*N* matching is performed without alignment steps.

Even for the same identification user, because of noises and local deformation, extracted minutiae exhibit different coordinates and angles over each acquisition. To solve this problem, an adaptive elastic matching algorithm in which tolerance levels are determined according to polar coordinates of minutiae was proposed in [2]. In this paper, the coordinate plane is divided into several fields according to the distance from the origin. Each field has its own level of tolerance for *x*- and *y*-coordinates. The first field has tolerance level of [-3, 3] which means errors between -3 and 3 in *x*-or *y*-coordinate are tolerated. Two transformed minutiae in the first field are considered to have matching coordinates if their coordinates do not differ more than this error range. Tolerance level for angles is 22.5 degree for all transformed minutiae.

In candidate list generation stage, all transformed minutiae set in identification table are compared with all transformed minutiae set in fingerprint database. A list with most similar enrollment users, *i.e.,* enrollment users with most number of matching transformed minutiae pairs, is generated

3 Implementation of Fingerprint Retrieval

To implement a real-time fingerprint retrieval system, we did hardware/software codesign of the fingerprint retrieval system. Fingerprint retrieval system performs the enrollment and identification processes. Enrollment process is performed by a software module that consists of minutiae information acquisition, table generation, and fingerprint database update. Identification process is performed by both hardware and software modules that consists of minutiae information acquisition, table generation, matching(1:*N*), and candidate list generation. Since the computation time of matching is proportional to the size of the fingerprint database, it is performed by the hardware module. Although table generation and candidate list generation do not require much computation, they are performed by the hardware module for an efficient implementation of the fingerprint retrieval system.

Our fingerprint identification system consists of fingerprint database, CPU, and H/W Accelerators. Fingerprint database stores enrollment tables containing the information of enrollment users. CPU performs the software modules described above and controls H/W Accelerators. A H/W Accelerator consists of a PCI controller, an FPGA, and SDRAMs. The PCI controller interfaces between PC and H/W accelerator. The FPGA performs table generation, matching(1:*N*), and candidate list generation in the identification process. SDRAMs store the enrollment tables downloaded from fingerprint database.

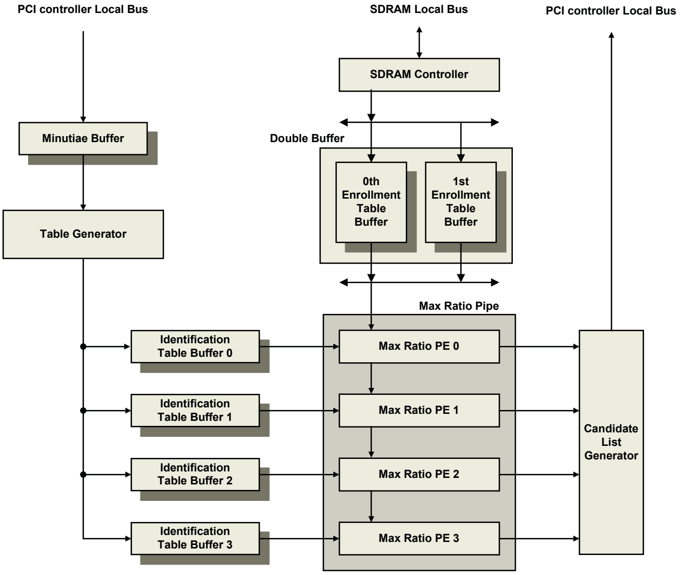


Fig. 1. Fingerprint Identification Hardware Module

Fingerprint identification hardware module consists of table generator, max ratio pipe, candidate list generator, SDRAM controller, and several types of buffers as shown in Fig. 1. Table generator, max ratio pipe, and candidate list generator perform table generation, matching(1:*N*), and candidate list generation, respectively. A minutiae buffer, four identification table buffers, and two enrollment table buffers are used to store data temporally.

Fig. 2 shows the architecture of Max ratio Processing Element(PE) in max ratio pipe, and it consists of transmit register, shift register, score PE, sum score, ratio, and max ratio. Transmit register transmits the identification table address and enrollment table data from the preceding max ratio PE to the following max ratio PE. Shift register shifts the identification table data to its right and stores it. Score PE, sum score, ratio, and max ratio compute similarity scores.

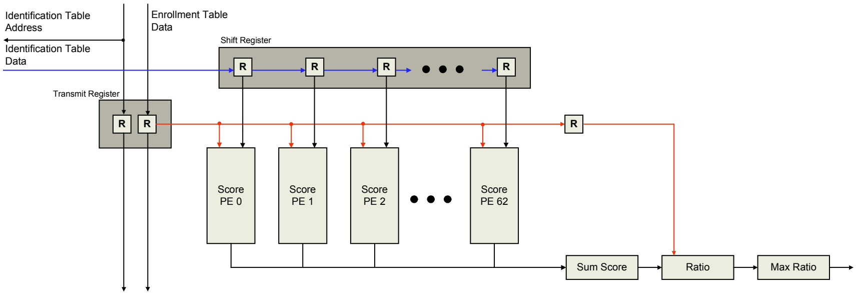


Fig. 2. The Architecture of Max Ratio PE

4 Experimental Results

In this paper, two PCI boards have been implemented to evaluate the performance of the proposed fingerprint retrieval system. The implemented PCI board contains a Xilinx FPGA with 4-million-gate and 2GB SDRAM. The total equivalent gate count for the design was 2,049,020. The PCI board and PCI bus are operated in 33 MHz.

The environment of Host PC where a PCI board is installed is summarized as follows: Intel Pentium 4 Zeon Dual Processor 1.7GHz, RDRAM 1,024MB, Intel i860 Chipset, 3×32bit, 2×64bit PCI Slot, and Linux operating system. To fit the hardware module into the FPGA, the number of minutiae per user is restricted up to 64, thereby limiting the hardware size. Therefore, a minutia of a user is transformed for up to 63 minutiae. The maximum number of transformed minutiae stored in an identification table is 64×63 = 4,032. If 1,024 users are enrolled, the fingerprint database has approximately 2^{22} transformed minutiae. Therefore, for the comparison between the identification table and the fingerprint database, the comparison of minutiae is executed $2^{22} \times 2^{12} = 2^{34}$ times.

To evaluate the retrieval ratio of the fingerprint retrieval system implemented on the FPGA-based hardware, the ETRI fingerprint database was used, and the size of captured fingerprint images is 248×292. To construct the database, a fingerprint is captured 4 times for a finger per user from the total 1,024 users. The time interval between two successive captures for a finger is longer than 30 minutes. For the sake of convenience, according to the capturing sequence(time interval), the captured fingerprint images are denoted as 4 different sets: A, B, C and D-set. For each experiment, the similarity between the queried fingerprint and the data of the same user is measured. For A-set, the percentage that the similarity of the same user is in similarity level 1~5 and 1~10 was 83.4% and 87.0%. The percentage for B-set was 86.2% and 89.2%, the percentage for C-set was 86.4% and 89.4%, and finally the percentage for D-set was 85.1% and 88.9%.

To evaluate the scalability of the implemented hardware for the fingerprint retrieval, we measured the execution time in two different cases: using one PCI board and using two PCI boards. The average execution time of the fingerprint identification hardware for 1,024 users and 2,048 users was 0.7 seconds and 1.4 seconds by using one PCI board, and the average execution time of the fingerprint identification hardware for 2,048 users was 0.72 seconds by using two PCI boards. We can see that the retrieval is performed in real time. Furthermore, in the hardware system with two

PCI boards, the execution time is not increased and kept steady, even if the number of users increases twice.

5 Conclusions

In this paper, a scalable hardware design for real-time fingerprint retrieval has been proposed and implemented on an FPGA-based platform. The straightforward approach to search a large fingerprint database is to scan the entire database and to compare the query against each reference model. Even though this approach could be implemented by hardware, the real-time requirement cannot be satisfied for large-scale databases. To solve this problem, we proposed a scalable solution by using PCI boards with FPGAs and SDRAMs, and provided parallelism both within the PCI board and between the PCI boards.

Acknowledgement

This research was supported by the MIC(Ministry of Information and Communication), Korea, under the Chung-Ang University HNRC-ITRC(Home Network Research Center) support program supervised by the IITA(Institute of Information Technology Assessment).

References

1. D. Maltoni, et al., *Handbook of Fingerprint Recognition*, Springer, 2003.
2. N. Ratha and A. Jain, "A Real-Time Matching System for Large Fingerprint Database," *IEEE Trans on PAMI*, Vol. 18, pp.799-813, 1996.
3. H. Wolfson and I. Rigoutsos, "Geometric Hashing: an Overview," *IEEE Computational Science and Engineering*, Vol. 4, pp. 10-21, Oct.-Dec. 1997.
4. C. Wang, V. Prasanna, H. Kim, and A. Khokhar, "Scalable Data Parallel Implementations of Object Recognition Using Geometric Hashing," *Journal of Parallel and Distributed Computing*, 21, pp. 96-109, 1994.
5. D. Warren, D. Kearney, and G. Wigley, "Field Programmable Technology Implementation of Target Recognition using Geometric Hashing," *Proc. of International Conference on Engineering of Reconfigurable Systems and Algorithms*, pp. 247-249, 2002.

Fast Video Retrieval via the Statistics of Motion Within the Regions-of-Interest

Jing-Fung Chen[1,2], Hong-Yuan Mark Liao[1], and Chia-Wen Lin[3,⋆]

[1] Institute of Information Science, Acadmeia Sinica, Taipei, Taiwan
[2] Department of Digital Media Center of National Taiwan Normal University, Taiwan
[3] Department of Computer Science and Information Engineering National,
Chung Cheng University, Taiwan

Abstract. It is a very important issue to quickly retrieve semantic information from a vast multimedia database. In this paper, we propose a statistic-based algorithm to retrieve the videos that contain the requested object motion from video database. In order to speed up our algorithm, we only utilize the local motion embedded in the region-of-interest as the query to retrieve data from MPEG bitstreams. Experimental results demonstrate that our fast video retrieval algorithm is powerful in terms of accuracy and efficiency.

1 Introduction

Due to the popularity of the Internet and the computing power developed in the past few years, efficient processing/retrieval of multimedia data has become an important issue. Among different types of media, video contains the most amounts of data and is relatively hard to be dealt with due to its complexity. In order to efficiently manage video data, a compact video representation scheme and an efficient retrieval algorithm are indispensable. A number of features have been proposed to represent the content of compressed video (MPEG 1/2) such as color, shape, motion, etc.. Many researchers use the motion vector as a feature to describe a video. Unlike color or texture, the motion feature can be extracted directly from the MPEG bitstreams. Therefore, there is no need to decompress a video and then extract features in the original domain. Conventional motion feature-based video retrieval algorithms [1–3] usually based their search strategy on the overall motion embedded in a video clip. However, in most cases only parts of the region in an image sequence are really "moving". Therefore, if the whole area of a frame is considered, some irrelevant regions in a frame which are actually not "moving" are also considered as useful features and used in the retrieval process. In this paper, we propose a region-of-interest (ROI) based video retrieval algorithm to solve the above mentioned problem. Using the proposed ROI-based approach, we are able to improve the efficiency as well as the accuracy of the retrieval results.

[⋆] This work is supported by the Ministry of Economic Affairs under contract no. 93-EC-17-A-02-S1-032

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 381–387, 2005.

In this paper, the preprocessing and the description of local motion will be introduced in Section 2 and Section 3, respectively. How to realize the concept of ROI and then incorporate it into our system will be described in Section 4. Then, we shall introduce the matching metric which can be used in the retrieval process in Section 5. Finally, experimental results will be reported and the conclusions will be drawn in Section 6 and Section 7, respectively.

2 Preprocessing

Since a shot is the most primitive unit with semantic meaning that can be used for video retrieval, a powerful shot change detection algorithm is indispensable. We use two powerful gradual shot change detection algorithms [5, 6] to divide video into shots. First, we use the camera model [4] to remove unreliable motion vectors and try to filter out the effect caused by camera motions. For a complete shot, it is possible to cover several GOPs (group of pictures). Under the circumstances, the original design of an MPEG bitstream cannot be used directly for computing the local motion (object motion) between every consecutive anchor frame pair (anchor frame means I- and P- frames). For the above mentioned discontinuity problem that does exist between two consecutive GOPs, we propose to use the motion vectors of the B-frame which is located right after the last P-frame in each GOP to solve the problem (shown in Fig. 1).

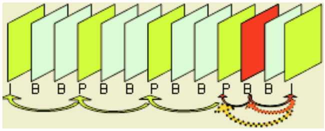


Fig. 1. Consecutive motion vector

3 Local Motion Extraction

In this study, we observed that the boundaries of a frame are relatively motionless and therefore should not be considered in the feature extraction process. In addition, the place in which the captions usually appear should also be discarded. Under these circumstances, we only use the central portion of a frame as valid area. Now, we are ready to discuss how to calculate the statistics of motion from a valid macro block sequence located in a shot. Using this statistics, we are able to conduct quantitative comparison between two distinct shots. The left hand side of Fig. 2 illustrates a typical shot consisting of n anchor frames. From this shot, we can derive $n-1$ local motions between any two consecutive anchor frames. Since a local motion vector derived from two consecutive macro blocks in a shot may be large in magnitude, we have to transform it into a smaller domain and, in the mean while, quantize it to facilitate the motion statistics calculation process. For a motion vector (x, y) located in the XY plane, we can transform it into the UV plane by the following equation:

$$\begin{aligned} u &= \lfloor x/I + 0.5 \rfloor \\ v &= \lfloor y/I + 0.5 \rfloor \end{aligned} \tag{1}$$

where I is an integer that can be used to control the degree of quantization. The quantization procedure is able to group the motion vectors that are close to each other into one bin.

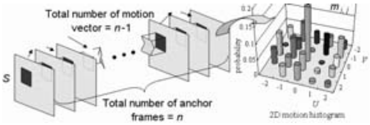


Fig. 2. An example showing how the motion vectors extracted from a macro block sequence are projected onto the UV plane

For calculating the statistics of motion from a valid macro block sequence (as indicated in Fig. 2), we have to do the following. First, let $m_{i,j}$ be the set of motion vectors of a valid macro block sequence. The first macro block of this macro block sequence begins from the i-th row, j-th column of the valid macro block region. The probability that the quantized (or transformed) motion vectors of this macro block sequence falls into the bin (u, v) can be calculated as follows:

$$p(LM = (u, v) | LM \in m_{i,j}) = \frac{\#\{LM | LM = (u, v)\}}{L} \tag{2}$$

where LM represents a motion vector after transformation (or quantization) and L is the total number of quantized motion vectors in this valid macro block sequence. The $\#\{LM | LM = (u, v)\}$ means the number of quantized motion vectors that fall into the (u, v) bin. Fig. 2 illustrates how to transform $n - 1$ motion vectors into the normalized probability distribution map located in the UV plane. In the example, the range of U and V are both from -2 to +2.

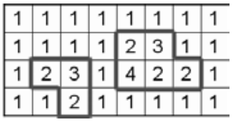
4 Regions of Interest

In this section, we shall describe how to extract the regions that contain the most amount of motion as the region-of-interest (ROI). Based on the formulation described in Section 3, the motion vectors of a macro block sequence can be projected onto the UV plane and then form a 2D motion histogram. It is reasonable to select those macro block sequences that have more non-zero bins as the members of the ROI. A macro block sequence that contains more non-zero bins means the activity inside this macro block in intensive.

In order to locate the ROI, we count the total number of non-zero bins from each valid macro block sequence as follows:

$$N^{\mathrm{ROI}}_{m_{i,j}} = \#\{\, p\ (LM = (u, v)\ |\ LM \in m_{i,j}) > 0\, \} \tag{3}$$

where $N^{\mathrm{ROI}}_{m_{i,j}}$ denotes the number of non-zero bins located in the macro block sequence. For those valid macro block sequences that with $N^{\mathrm{ROI}}_{m_{i,j}}$ value larger than 1, we consider them as the constituents of the ROI. Fig. 3 is an example showing the distribution of the number of non-zero bins with respect to a valid macro block region. The aggregation of the two regions shown in Fig. 3 will form the ROI in this case.



1	1	1	1	1	1	1	1
1	1	1	1	2	3	1	1
1	2	3	1	4	2	2	1
1	1	2	1	1	1	1	1

Fig. 3. An example showing how the ROI is formed

5 Matching Procedures

For comparing two distinct shots, the comparison is of the form of comparing two probability distribution functions. Here, we shall use the so-called Bhattacharyya distance [7] to do the job. The Bhattacharyya distance is a well-known metric which is defined for measurement of the correlation between two arbitrary statistical distributions. For any two arbitrary distributions $p(x|\omega_1)$ and $p(x|\omega_2)$ of classes ω_1 and ω_2, respectively, the continuous form of the Bhattacharyya distance is defined as [7]:

$$D(\omega_1,\omega_2) = -\ln \int (p(\mathbf{x}|\omega_1)p(\mathbf{x}|\omega_2))^{1/2} d\mathbf{x} \tag{4}$$

However, in the previous section we have described that the distribution of motion vectors in a shot is formulated as the discrete format. Therefore, we need a discrete Bhattacharyya distance to perform the shot comparison task. Let $m_{i,j}$ and $m'_{i,j}$ be the set of motion vectors extracted from the valid macro block sequence at (i,j) location of two distinct shots (the two shots may have different lengths). Then, based on the definition made in Eq. (4), we can define a discrete Bhattacharyya distance for the above two distinct macro block sequences as follows:

$$\begin{aligned} d(m_{i,j}, m'_{i,j}) = \sum_{u,v} (p(LM = (u,v)|LM \in m_{i,j}) \\ p(LM' = (u,v)|LM' \in m'_{i,j}))^{1/2} \end{aligned} \tag{5}$$

For calculating the distance between two shots, we use the example shown in Fig. 4 to explain how the Bhattacharyya distance works. Fig. 4 illustrates two shots having different lengths. The number of anchor frames for the upper shot and the bottom shot are n and p, respectively ($n \neq p$). Taking a valid macro block sequence at the same position from the two different shots, we shall have $n-1$ and $p-1$ motion vectors, respectively, extracted from the upper shot and the bottom shot.

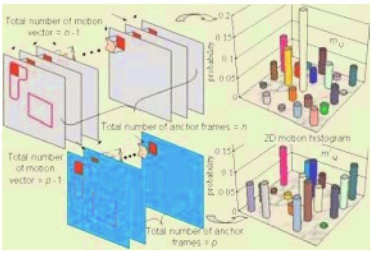

Total number of motion vector = n - 1
0.2
0.15
0.1
0.05
0
probability
m_{uv}
Total number of anchor frames = n
2D motion histogram
Total number of motion vector = p - 1
0.15
0.1
0.05
0
probability
m'_{uv}
Total number of anchor frames = p

Fig. 4. An example showing how the motion vectors that belong to two macro block sequences of two distinct shots are extracted and mapped onto the UV plane. The thick-line at first frame of video sequence show the ROI. The correlation between the two macro block sequences is calculated by multiplying bin-to-bin probabilities and then summing them up

After transformation (quantization), we are able to quantize both motion vector sets onto the UV plane as indicated on the right hand side of Fig. 4. The comparison using the Bhattacharyya distance shown in Eq. (5) can be carried out as follows. Eq. (5) is the Bhattacharyya distance designed to measure the similarity between two macro block sequences located at the same position but from different shots. In order to calculate the overall Bhattacharyya distance of the ROI between two arbitrary shots, one has to accumulate the measured distance from all macro block sequence pairs located in the ROI. The equation for calculating the overall similarity, $D(S, S')$ is as follows:

$$D(S, S') = -\ln\left(\frac{\sum_{\text{ROI}} d(m_{i,j}, m'_{i,j})}{N}\right) \qquad (6)$$

where S and S' represent two distinct shots, and N represents the total number of valid macro block sequences existing in the ROI of a shot.

6 Experimental Results

In order to show the effectiveness of the proposed method, we have tested our algorithm against a 1682 - shot video database. First, we used the gradual shot change detection algorithms proposed in [5, 6] to extract 1682 complete shots from six digital videos. For each constituent shot in the database, we used the same method as described in this paper to calculate its statistics of motion (off-line). The length of the six digital videos were 55 minutes (503 shots, documentary, video #1), 52 minutes (405 shots, documentary, video #2), 29 minutes (241 shots, commercial, video #3), 38 minutes (193 shots, news, video #4), 38 minutes (283 shots, sports news, video #5), 17 minutes (57 shots, home video, video #6), respectively. The reason why we chose these videos was due to their variety. Fig. 5 and Fig. 6 show two retrieval results from our experimental results. In Fig. 5, we randomly chose a video shot which was one of the database shots as the query and retrieved the top five hits from the database. Fig. 6 shows

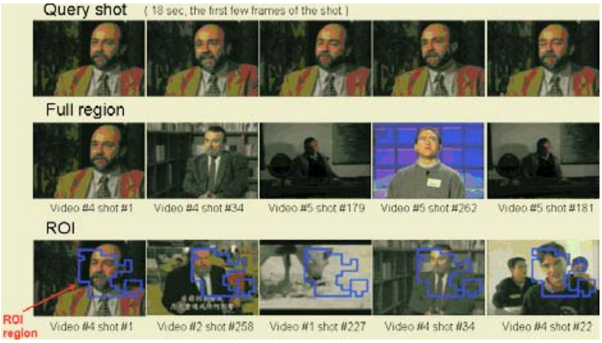


Fig. 5. Retrieved results of query shot (video #4 shot #1)

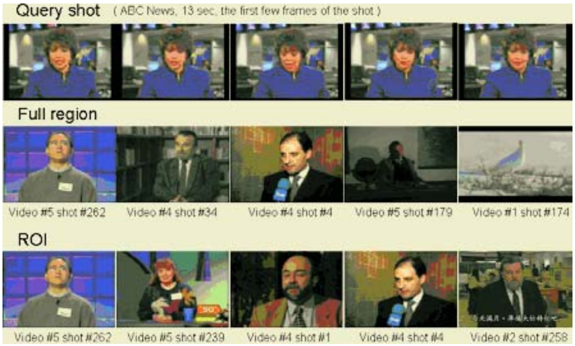


Fig. 6. Retrieved results of outside testing query shot

another retrieval results that used another video shot which was not one of the members of the database as the query.

In order to show that the ROI strategy is indeed useful in improving the retrieval outcome, we also conducted the same experiments by using the all-region motion as the query. Table 1 illustrates how the ROI concept improves the efficiency of the algorithm.

Table 1. The computation reduction rate for inside query and outside query with ROI-based method in comparison to Full-region-based method

	Computation Reduction Rate
Query I	65%
Query II(ABC news)	65.6%

7 Conclusions

We have proposed a fast video retrieval algorithm which could retrieve shot efficiently and accurately. The proposed approach used the statistics of motion extracted from a shot as the search features. By using automatic selection ROI to query shots, the procedure also can reduce the interference from other non-ROI macro block sequences to increase performance of query results. And the algorithm returned the accurate result in mini-seconds. Experimental results have demonstrated that the proposed fast video retrieval algorithm is indeed powerful.

References

1. Peker, K.A.and Divakaran, A., "A novel pair-wise comparison based analytical framework for automatic measurement of intensity of motion activity of video segments," *Proc. ICME*, pp. 729-732, Aug. 2001.
2. Chen, J.F., Liao, Mark H.Y. and Lin, C.W., "Fast video retrieval via the statistics of motion," *Proc. ICASSP*, Mar. 2005.
3. Sun, X., Divakaran, A. and Manjunath, B.S., "A Motion Activity Descriptor and Its Extraction in Compressed Domain," *Proc. PCM, Bejing, China*, Oct. 2001.
4. Srinivasan, M.V., Venkatesh, S. and Hosie, R., "Qualitative Estimation of Camera Motion Parameters form Video Sequences," *Pattern Recognition*, Vol. 30, No. 4, pp. 593-606, 1997.
5. Shih, C.C., Tyan, H.R. and Liao, Mark H.Y., "Shot Change Detection based on the Reynolds Transport Theorem," *Proc. PCM, Beijing, China*, Vol. 2195, pp. 819-824, Oct. 2001.
6. Su, C.W., Tyan, H.R., Liao, Mark H.Y. and Chen, L.H., "A Motion-tolerant Dissolve Detection Algorithm," *Proc. ICME, Lausanne, Switzerland*, Aug. 2002.
7. Chen, L.F., Liao, Mark H.Y., Lin, J.C. and, Han, C.C., "Why Recognition in a Statistics-Based Face Recognition System should be Based on the Pure Face Portion: a Probabilistic Decision-Based Proof," *Pattern Recognition*, Vol. 34, No. 5, pp. 1393-1403, 2001.

Information Theoretic Metrics in Shot Boundary Detection

Wengang Cheng, De Xu, Yiwei Jiang, and Congyan Lang

Department of Computer Science, Beijing Jiaotong Univ., 100044 Beijing, P.R. China
wengangcheng@163.com

Abstract. A favorable difference metric is crucial to the shot boundary detection (SBD) performance. In this paper, we propose a new set of metrics, information theoretic metrics, to quantitatively measure the changes between frames. It includes image entropy difference, joint entropy, conditional entropy, mutual information and divergence. They all can be used to cut detection. Specially, the image entropy and joint entropy are good clues to fade detection, while mutual information, joint entropy and conditional entropy are less sensitive to illumination variations. The theoretic analysis and experimental results show that they are useful in SBD.

1 Introduction

SBD servers as the preliminary step to video content analysis. The detection of shot boundaries is dependent on the fact that consecutive frames on either side of a boundary generally display a significant change in content. Usually, a difference measure between successive images is defined first. If the difference exceeds a given threshold, it indicates a shot boundary. Hence, establishing suitable metrics is the fundamental issue in SBD. The ideal metric will be able to differentiate between shot transitions and other image changes, such as motion, luminosity change, noise, etc.

Different metrics used in SBD can divided into two major types: (1) those based on local pixel feature comparison, such as pixel values [1] and edges [2], and (2) those based on global features such as histograms and statistic distributions of pixel-to-pixel change. So far, the most popular metric is histogram-based frame difference, which achieves good tradeoff between accuracy and speed [3].

As a further research on our former work [4], this paper analyses the SBD in a new viewpoint, information theory. We present a new set of metrics called information theoretic metrics, to quantitatively measure the changes between consecutive frames. Some useful properties of these metrics are analyzed. Both the theoretic analysis and the experimental result verify they are useful in SBD.

The paper is organized as follows. Section 2 presents the proposed information theoretic metrics. Based on the proposed metrics, we give some experimental results in Section 3. Finally, Section 4 concludes the paper.

2 Information Theoretic Metrics

As a media, video servers as the information carrier. Each frame can send certain information to the audience. Therefore, analysis frames from the view of information

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 388–394, 2005.

is nature and straight. Here, we will model the characters of information changes within the same shot and at the boundaries of a shot to find usable clues for SBD.

2.1 Metrics

2.1.1 Image Entropy Difference

If we take a frame as a discrete information source, a video sequence can be regarded as a series of information sources. As we know, the frames of the same shot have similar visual contents. Considering the temporal and dynamic property of video, these information sources, in form of frames, have similar information source space. As a result, the information quantity will not change too much within a shot. On the contrary, when there is a shot transition, the information quantity will change dramatically. Thus the quantity of information provided by a frame is a good feature for SBD.

In information theory, entropy is a measure of the average amount of information we obtain by receiving one symbol from the information source. Here, image entropy (IE) can measure the information contained in a frame. The difference between the IE of video sequences becomes beneficial to SBD. The entropy of a discrete random variable (DRV) X with probability mass function $p(x)$ is defined as:

$$H(X) = -\sum_{x} p(x) \log p(x) \ . \tag{1}$$

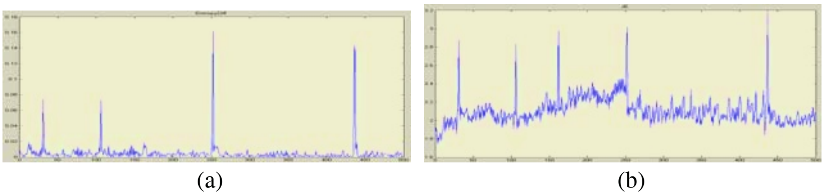
The IE $H(F_i)$ can be calculated with the normalized frame histogram, and the image entropy difference (IED) is the absolute value between two IEs. If IED arises abruptly, it indicates there is a shot transition, as Fig. 1(a) illustrates.

2.1.2 Joint Entropy

When considering two DRVs X and Y at the same time, it is possible to measure the degree of uncertainty or information associated with them. It is called the joint entropy (JE), $H(X, Y)$. If X and Y are jointly distributed according to $p(x, y)$, the $H(X, Y)$ is:

$$H(X,Y) = -\sum_{x}\sum_{y} P(x,y) \log p(x,y) \ . \tag{2}$$

JE measures how much entropy is contained in a joint system of two DRVs. Its value lies in the range of [max($H(X)$, $H(Y)$), $H(X)$+$H(Y)$]. If X, Y independent, it takes maximal value. Consider a video sequence, when the adjacent two frames are visual similar, in other words, they have strong correlation in visual contents, the JE will be relatively small. While there is a shot transition, the JE reaches the local maxima. In fact, the JE of two frames increases with their decreasing visual similarity. So, inter-frame difference can be measured by JE of adjacent frames, Fig. 1(b) is an example.



(a) (b)

Fig. 1. IED and JE for SBD: (a) IED for a video segment with four cuts; (b) JE for a video segment with five cuts

2.1.3 Conditional Entropy

Let *X* and *Y* be two DRVs with joint distribution *p*(*x*, *y*) and conditional distributions *p*(*x*/*y*), the conditional entropy (CE) of *X* and *Y* is:

$$H(Y/X) = -\sum_{x,y} p(x,y)\log p(y/x) \tag{3}$$

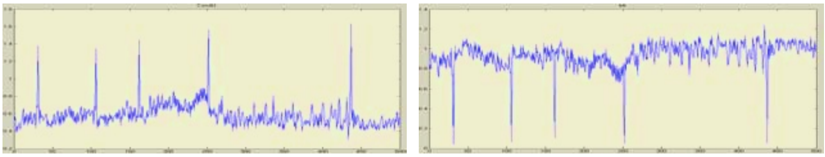
CE measures how much entropy a DRV has remaining if we have already learned completely the value of a second DRV. Given two visual similar frames, we cannot get much information from the second frame when we have known the first one. In this example, the CE of the two frames is small. On the contrary, given two frames different in contents, it is evident large CE of the two frames can be gotten here. Corresponding to above two cases, consecutive frames within the same shot or at shot boundaries show such visual characters. Fig. 2(a) gives an example.

2.1.4 Mutual Information

Let *X* and *Y* be two d.r.v.s with marginal probability distributions *p*(*x*) and *p*(*y*), and probability distribution *p*(*x*,*y*), then the mutual information (MI) between *X* and *Y* is:

$$I(X,Y) = \sum_{x,y} p(x,y)\cdot\log(\tfrac{p(x,y)}{p(x)\cdot p(y)}) \ . \tag{4}$$

MI measures the information that one DRV contains about a second DRV. A large MI between two adjacent frames means they have a large dependency, while a low value shows low dependency and therefore a shot boundary. It has been used in [4,5,6] for SBD. Fig 2(b) gives an example.



(a) (b)

Fig. 2. CE and MI for a video segment with five cuts: (a) CE; (b) MI

2.1.5 Divergence

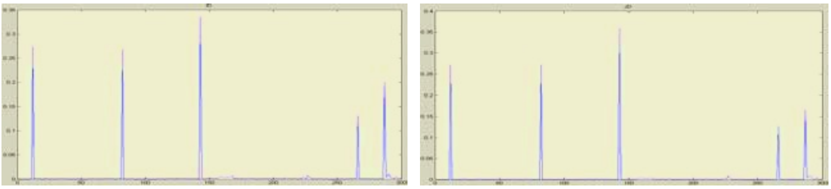
Let *X* be a DRV, *P* and *Q* be two probability distributions of *X*. The *I* directed divergence (ID), also called relative entropy or Kullback-Leibler's distance, is defined as:

$$ID(P,Q) = -\sum_{x} P(x)\log\left(\tfrac{P(x)}{Q(x)}\right) \ . \tag{5}$$

ID is a difference measure between two probability distributions. Since each frame takes a probability distribution, the divergence between two probability distributions of consecutive frames gives the difference of information quantity. ID is non-negative, additive but not symmetric. To obtain a symmetric measure, one can define:

$$JD(P,Q) = ID(P,Q) + ID(Q,P) = \sum_{x}\left(P(x)-Q(x)\right)\log\left(\tfrac{P(x)}{Q(x)}\right) \ . \tag{6}$$

which is called *J* divergence (JD). Clearly, ID and JD share the same function as dissimilarity between two frames, as Fig. 3 shows.



(a) (b)

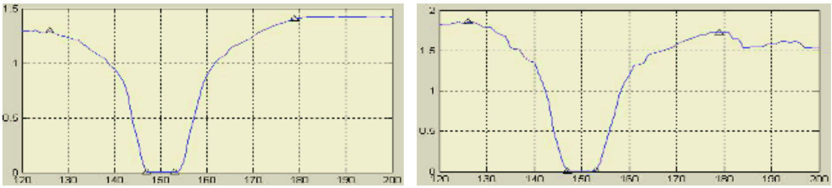
Fig. 3. Divergence for a video segment with five cuts: (a) ID; (b) JD

2.2 Characters

2.2.1 Fade Detection

From the view of information, the information quality continues increasing during fade in, along with the frame contents appears gradually. From the view of histogram, the histogram shows increasing dispersion during fade in. Since entropy has the abilities to measure not only the information quality, but also the dispersion of histogram, IE can describe fade in very well. To consider the information quality and the joint histogram of consecutive frames, we can get that JE follows the same changing pattern as IE. Fig.4 is the curve of IE and JE during a fade out and a fade in.

It is very easy to locate the starting point of fade in, where IE is near zero. To seek the ending point, we can utilize the monotonic property of the entropy values. In fact, the inflection point of the entropy curve is the ending point of fade in. Often the entropy value is not strictly monotonic during fade in/out, we can resort to a tolerant value or the smoothed entropy values, so as to the starting point of fade out.



(a) (b)

Fig. 4. IE and JE for fade detection: (a) IE; (b) JE

2.2.2 Illumination Variation

Illumination variation (IV) has bad influences in most of the existing methods. For example, histogram-based methods are very sensitive to IV. More than metrics merely, MI, JE and CE show good performance under this condition. These metrics, which take into account two frames at the same time, are computed from joint histogram and/or conditional histogram. Different with histogram of a single frame, the joint histogram and conditional histogram smooth the inter-frame difference to some extent actually. Therefore, they are relatively, although not completely, non-sensitive to IV. Fig. 6 shows that histogram-based methods are easy to find error boundaries when there are dramatic illumination changes, while the others are much better.

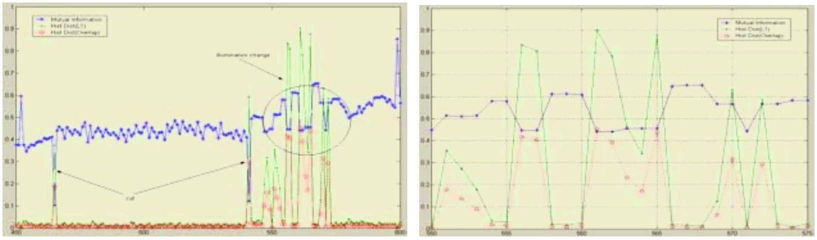


Fig. 5. Curves of three metrics for a video segment with IV.(a) MI, histogram L_1 distance and histogram intersection values; (b) the zoomed-in region containing dramatic IV

3 Experimental Results

To test the validity and evaluate the performance of these metrics, we conduct three groups of experiments, each for a special objective. Both the recall and precision are used to evaluate the results. In these experiments, a sliding widow scheme is used to get a self-adaptive threshold locally.

As described in Section 1, the histogram-based method can achieve high performance, so we include it in our experiments and two histograms are matched using the L_1 distance. All the metrics are calculated using the gray values (equally divided into 64 levels) and the same bins of histogram (64) on the whole frame(not block-based). Shot boundary is declared when the inter-frame difference exceeds the adaptive threshold.

Group A is used to test the performance of cut detection. Experimental dataset consists of several sequences selected from TRECVID-2001 and TRECVID-2002. The chosen videos are edited firstly, and there is only one type of shot transitions (i.e. cut) left in these segments. All the metrics in Section 2.1 and histogram L_1 distance(L_1 Dist) are considered. Table 1 gives the results. There is a pair of real values *a*, *b* in each table cells, which means the precision *a* and the recall *b*.

Table 1. Detection performance for cuts

Name	IED	JE	MI	CE	ID	L_1 Dist
Cut1.mpg	0.662, 1.0	0.832, 0.840	0.983, 0.888	0.945, 0.827	1.0, 0.953	1.0, 0.961
Cut2.mpg	0.793, 0.950	0.824, 0.760	0.925, 0.835	0.833, 0.782	0.940, 0.915	0.919, 0.883
Cut3.mpg	0.791, 0.950	0.872, 0.897	1.0, 0.980	0.938, 0.901	1.0, 1.0	1.0, 1.0
Cut4.mpg	0.631, 0.802	0.755, 0.810	0.940, 0.913	0.916, 0.826	0.965, 0.923	0.970, 0.910

They all work well for cut detection. In ID-based method, peak values at boundaries are much higher than the values obtained from frames without shot changes. ID gives the best results, while IED gets relatively worse results for some shots having unchanging visual primitives. As nearly the same results returned by ID and JD, we just give one group of results provided by ID, so as to the two conditional entropy values.

Fade detection performance is evaluated in *Group B*. We select three videos, which contain a lot of fade in/out effects, from TRECVID-2002. Then, they are edited by cutting away other types gradual transitions (e.g., dissolve). The edited video segments are used as the dataset. Boreczky [3] found twin-comparison method is an efficient algorithm. Although the twin-comparison method cannot recognize different types of gradual transition, it does not influence our test because there is only one type of gradual transition in our dataset. IE, JE and the histogram-based twin-comparison method are used for fade detection. Table 2 is the results.

Table 2. Detection performance for fades

Name	IE	JE	Twin-Comparison
Fade24.mpg	1.0, 0.958	1.0, 0.916	0.833, 0.875
Fade20.mpg	1.0, 0.950	1.0, 0.900	0.900, 0.850
Fade17.mpg	1.0, 0.825	1.0, 0.825	0.882, 0.823

No missed fade in/out found in the method based on IE or JE. However, monochrome frame is a potential factor to false alarm. It is evident that the methods based on the information metrics give better results than twin-comparison method.

Group C aims to compare the metrics in the aspect of robustness to IV. Two segments selected from horror movies serve as our dataset. Each segment contains cuts and dramatic IV. MI, JE, CE and L_1 Dist are used here. Table 3 presents the results.

Table 3. Detection Performance under Illumination Variations

Name	MI	JE	CE	L_1 Dist
Light20.mpg	1.0, 0.833	0.909,0.800	0.952, 0.714	1.0, 0.571
Light43.mpg	0.935, 0.914	0.914,0.860	0.895, 0.860	0.977, 0.716
Light11.mpg	1.0, 0.846	0.833, 0.714	0.714,0.666	1.0, 0.550

Although L_1 Dist gets high recall, the precision is the worst one, for it mistakes the illumination changes as cuts. On the contrary, because ME, JE and CE are not so sensitive to light changes, they achieve better results.

Although all the proposed metrics except IED cannot fully satisfy the conditions of distance function (positivity, symmetry and the triangle inequality), and are therefore not true metrics, they are capable of measuring the inter-frame similarity or dissimilarity very well. In this paper, the information theoretic metrics are calculated using the histogram. As a result, some good qualities of histogram are kept, while the computation complexity increases on the other hand.

4 Conclusions

In this paper, we analyse the SBD in a new viewpoint, information theory. A new set of metrics, information theoretic metrics, is put forward. Experimental results show the validities of these metrics. They all can be used in cut detection. The IE and JE show prominent properties in fade detection. MI, JE and CE are not sensitive to IV.

References

1. Shahraray, B.: Scene change detection and content-based sampling of video sequences. In:IS&T/SPIE'95 Digital Video Compression: Algorithm and Technologies, Vol.2419, (1995) 2-13
2. Zabih,R., Mai,K., Miller,J.: A robust method for detecting cuts and dissolves in video sequences. In: Proc. of ACM Multimedia (1995)
3. Boreczky, J.S., Rowe L.A.: Comparison of video shot boundary detection techniques. In: Sethi, K., Jain, R.C., Ishwar, K.S., (eds.). Proceedings of the SPIE Conference on Storage and Retrieval for Still Images and Video Databases IV., SPIE Press,Vol. 2664 (1996) 170-179
4. Cheng, W.G., Liu Y.M.: Shot boundary detection using the knowledge of information theory. In: IEEE ICNNSP, (2003) 1237-1241
5. Butz, T., Thiran, J.P.: Shot boundary detection with mutual information. In: IEEE ICIP, (2001) 422-425
6. Cernerkova,Z., Nikou,C., Pitas,I.: Shot detection in video sequences using entropy-based metrics. In: IEEE ICIP, (2002) 421-424

Design of a Digital Forensics Image Mining System

Ross Brown[1], Binh Pham[1], and Olivier de Vel[2]

[1] Faculty of Information Technology, Queensland University of Technology,
GPO Box 2434, Brisbane 4001, Australia.
{r.brown,b.pham}@qut.edu.au
[2] Information Networks Division, Defence Science and Technology Organisation,
PO Box 1500, Edinburgh 5111, Australia
Olivier.DeVel@dsto.defence.gov.au

Abstract. Increasing amount of illicit image data transmitted via the internet has triggered the need to develop effective image mining systems for digital forensics purposes. This paper discusses the requirements of digital image forensics which underpin the design of our forensic image mining system. This system can be trained by a hierarchical Support Vector Machine (SVM) to detect objects and scenes which are made up of components under spatial or non-spatial constraints. Forensic investigators can communicate with the system via a grammar which allows object description for training, searching, querying and relevance feedback. In addition, we propose to use a Bayesian networks approach to deal with information uncertainties which are inherent in forensic work. These inference networks will be constructed to model probability interactions between beliefs, adapt to different users' retrieval patterns, and mimic human judgement of semantic content of image patches. An analysis of the performance of the first prototype of the system is also provided.

1 Introduction

Digital forensics is the application of computer analysis techniques to determine potential legal evidence of computer crimes or misuse that are caused by unauthorised users or by unauthorised activities generated by authorised users. The significance of digital forensics can be seen from the 2003 Computer Crime and Security Survey, published jointly by the Computer Security Institute and FBI, which reported total annual losses incurred by unauthorized computer use exceeding USD200 million for 251 organizations surveyed [1]. Digital forensics covers a wide range of applications such as law enforcement, fraud investigation, theft or destruction of intellectual property. Techniques used for such investigations are varied and may include data mining and analysis, timeline correlation, information hiding analysis, etc. Since multimedia format is widely used and readily available via the Internet, there are increasing criminal activities in the last few years, which involve the transmission and usage of inappropriate material such as child pornography in this format. Hence, much forensic evidence comes in the form of images or videos that contain objects and/or scenes that may be related to criminal behaviours. A typical investigation in digital forensics can generate large image and video data sets. For example, a disk can easily store several thousands of images and videos in normal files, browser cache files and unallocated space (i.e., non-file system areas on the disk which may contain fragments of

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 395–404, 2005.

files). It has been estimated that, as of late 2003, there exist some 260 million pages of pornography on the Internet [2]. This can make the task of searching for, and retrieving, images/videos very time consuming. Digital Image Forensics (DIF) efficiently seeks for evidence by using appropriate techniques based on image analysis, retrieval and mining. Owing to rising criminal activities via the internet, the use of such techniques for investigative purposes have only recently emerged, although they have been intensively researched over the last three decades for many other important applications: medical diagnosis, mineral exploration, environmental monitoring and planning, aerial surveillance, etc.

Content-based approaches have been developed that are based on some general low-level visual features such as colour, shape, texture e.g. [3]. Search-by-example is a common practice whereby an image is supplied and the system returns images that have features similar to those of the supplied image. The similarity of images is determined by the values of similarity measures that are specifically defined for each feature according to their physical meaning. Since the quality of the retrieval results relies on the choice of features and their similarity measures, much research has been focused on identifying features with strong discriminatory power and similarity measures that are meaningful and useful. In addition, we would ideally want a more "intelligent" system which can include high-level knowledge, deal with incomplete and/or uncertain information, and learn from previous experience. Such systems could include, for example [4]:

- Model-based Methods: A model of each object to be recognised is developed. These objects are classified using their constituent components that in turn are characterised in terms of their primitives,
- Statistical Modeling Methods: Statistical techniques are used to assign semantic classes to different regions/objects of an image, and
- User Relevance Feedback Methods: User feedback is required to drive and refine the retrieval process. The system is thus able to derive improved rules from the feedback and consequently generate better semantic classes of images.

Model-based methods exploit detailed knowledge about the object and are capable of reasoning about the nature of the object. However, the models created are often handcrafted and cannot easily improve their performance by learning. Statistical modelling techniques rely on statistical associations between image semantics and, as such, do not require the generation of any complex object model. Such associations can be learned using the statistical model. However, it is difficult for the investigator to interpret some of the results (e.g., "why are these objects in the image scene similar?") because statistical modelling techniques cannot easily reason with any high-level knowledge about the regions and image scene. User relevance feedback techniques inherently capture continuous learning as the system is able to build up a knowledge base of past user feedback. Quite elaborate feedback mechanisms can be implemented, e.g., ranking of images, input from collaborating investigators etc. (e.g., [5]). Image mining in digital forensics would ideally use a combination or hybridization of these methods.

In Section 2, we discuss various requirements of image forensics in terms of types of search, level of performance, learning ability and user interfaces. Section 3 presents the operation model of our forensic image mining system and the motivations behind its design. Section 4 gives an overview of how an SVM is used for training to detect objects and scenes which are described as a hierarchy of components and constraints, while Section 5 briefly describes the grammar which supports the modes of interaction between users and the system for specification, querying and relevance feedback for continual improvement. A summary of performance analysis of the prototype implemented so far is also provided. More details can be found in our two previous papers [6, 7].

2 Requirements of Forensic Image Mining

Image mining is only one of many different activities undertaken during a digital forensic investigation. As mentioned previously, a digital forensic investigation can involve a large number of data/evidence derived from a variety of sources as, for example: structured and unstructured files (e.g., text, marked-up text, databases), images, videos, music, network packets and router tables, process tables, telephone call records and so on. Also, an investigation may involve access to partial data (such as disk clusters), hidden data (e.g., data in disk partition gaps, steganography), encrypted data etc. The basic process in an investigation involving digital evidence would consist of a sequence of rigorous steps, including: extracting all of the data whilst maintaining the integrity of the original media and ensuring the chain of custody, filtering out the irrelevant data and identifying the useful data and metadata (e.g., file timestamps), deriving timelines, establishing the relationships between the disparate data (link analysis and link discovery), establishing causal relationships (causal analysis), identifying and extracting profiles, generating a comprehensive report etc.

The challenge in digital forensics is to find and discover forensically interesting, suspicious, or useful patterns or partial patterns in the potentially very large (now of the order of terabytes, TB) data sets. This task is analogous to the "needle-in-the-haystack" problem or, in the case of partial patterns located in multiple sources of evidence, "bits-of-needles-in-bits-of-haystacks". Furthermore, digital forensics has some unique requirements that make it rather different from traditional pattern extraction activities, for example [4]:

- Digital forensics deals with data instances that are both unrelated and related. That is, data instances may have multiple relations (e.g. networks of computer users, email cliques, geographical co-location etc.).
- The "interestingness" of data or sequences of events may be determined their low frequency of occurrence and possibly their non-repetitiveness. Unusual events may be more relevant in an investigation (i.e. we may be interested in the 'outliers').
- Sources of data in digital forensics are large, thereby requiring the consolidation of multiple data sources.
- Data sources may be high-dimensional and involve very different and sparse attributes.
- False negatives need to be minimised as the cost of "missing the needle in the haystack" is large. On the other hand, the number of false positives is not an overly sensitive parameter though, clearly, it should be kept to a minimum.

3 Operational Model

In order to design and implement an efficient image mining system architecture, an operational model of the digital forensic image mining process was developed. This model reflects the procedures undertaken by an investigator during a typical digital forensics investigation.

The model consists of two “activities”, namely one involving the rapid reduction of the large quantity of evidence that is involved in a case, and one involving the core image mining activities that deal with the actual image retrieval process for digital forensic examination. The former activity, as mentioned in Section II, involves the execution of a chain (in reality, a forest of connected trees) of forensic tools for analyzing the content of large data streams (disks and other data), filtering the data streams for data reduction, extracting meta-data (eg, file timestamps) etc. to downstream analysis and decision making that leads to a successful investigation. The latter activity is simply one of the many possible forensic tools deployed in the case investigation graph. The core image mining operational model follows two stages, namely the training phase and the testing or classification phase.

The training phase, also referred to as the classification model-generation phase, builds the object models relevant to the particular domain at hand. This phase is usually undertaken by an experienced investigator who has an insight into the object types involved in the particular case under investigation, an understanding of the classifier, knowledge of the object layout (eg, constraints such as positions, orientations etc.) and so on. The investigator will also be responsible for providing the relevance feedback on *a priori* evidence (eg, images from similar cases) in order to refine and improve the quality of the classification model. We propose to use a Bayesian Network for query refinement with a set of relevance feedback parameters (see Section VI). The testing phase uses the refined classification model (given by the set of model parameters) developed during the training phase to classify the set of images found in the case under investigation.

We have designed and developed a complete operational system for digital forensics which implements both the digital forensic examination process (the chain of forensic tools) as well as a prototype model-building and classifier system that focuses on the core image mining component of the operational model. The digital forensics investigative system (called “CFIT”, or computer forensics investigative toolkit) is not described here. In this paper, we focus on the model-building and classifier system.

4 Detection of Component-Based Objects and Scenes

There have been various component-based systems which deal with human detection. For example, features such as eye, nose, and mouth are first detected and then combined in a spatially constraint configuration in order to determine a face e.g. [8]. Other systems detect humans and their actions for various purposes: surveillance (e.g. detection of criminal activities [9]; movement recognition (e.g. gesture recognition for interactive dance systems [10]. The underlying models for such methods can be grouped in two main categories: task-specific models and general models that can be applied to specific tasks. The task-specific approach constructs a model from the components of a human silhouette and tightly coupled it with constraints that govern a specific action of interest e.g. [10], [11], and [12]. This approach is rather restrictive and does not provide a framework that can be readily extended in order to model different behaviours for other applications. The general approach, on the other hand, constructs a model from primitives in a bottom-up fashion and uses a regular grammar to represent various modes of motion and interactions e.g. [13], [14], and [15]. The system is then trained using models that represent certain exemplar behaviours. A special type of statistical models called Hidden Markov Models (HMM) [16] is used to represent both *a priori* knowledge and new knowledge resulted from new behaviours. Low level primitives are firstly detected before they are passed into the grammar for behaviour analysis. These systems, although robust, rely on motion information to resolve ambiguities.

We extend the approach by [17] which used Haar wavelet coefficients as features and SVMs (Support Vector Machine) for training. In their system, the magnitude of the coefficients of two scales (16x16 and 8x8 pixels) and three orientations (horizontal, vertical, diagonal) that indicate the intensity variation are used to locate the position of the components of objects. This multi-level approach is robust and flexible for object configuration design. One drawback is that difficulties due to image scaling and transformations have not been addressed. Our image mining system for computer forensic purposes allows the use of other features (e.g. texture features) in addition to Haar coefficients. We also investigate the effects of using different colour spaces, and of image scaling and transformations. In addition, we examine the needs of effective communication and usage of the system by forensic investigators and relevance feedback for continuous improvement. To this end, we develop a grammar to facilitate the specification of objects, scenes and their relationships. This grammar can also help to filter out invalid configurations. Relevance feedback will be provided via a Bayesian inference network [18].

The image mining module consists of two main parts: training and querying. We separate the two processes because of the differences in technical proficiency and forensic expertise required by each operation. The model trainer sets up parameters used by the classifier and constraints placed on the components of the model in order to train the SVM to recognize certain patches of an image. The query operator runs a query for the classification of a given image, using previously set up queries. Fig. 1 shows the relationships between these two processes.

The training process firstly segments the images in the training set, then calculates feature parameters and obtains appropriate constraints on the model components. These are stored in a database of scene descriptions. A bootstrapping process is then performed until the results are acceptable. This process involves the tweaking of parameters relating to features and constraints, and the retraining of patch detectors after false positive and false negative images from the test runs are added. The output models are stored in an object model database to be used later as query models by the query operator (Fig. 2). In the querying process, the operator supplies an unclassified image. The system segments the image to obtain feature vectors of image patches, then compares them with the models in the object model database and the scene descriptions to obtain a classified image (Fig. 3).

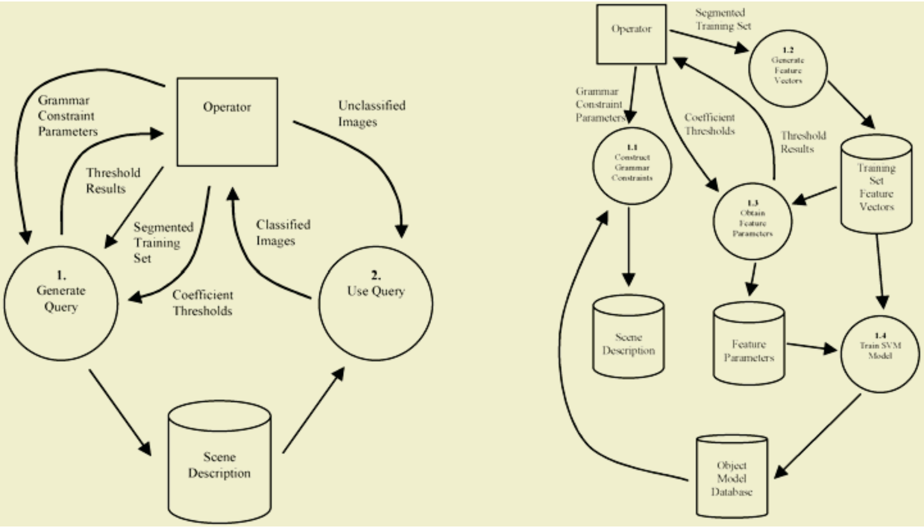


Fig. 1. Overview of the training and querying processes (left) and detail of query generation process (right)

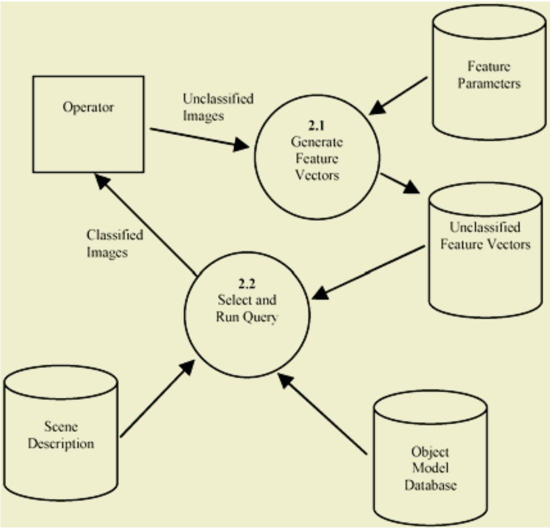
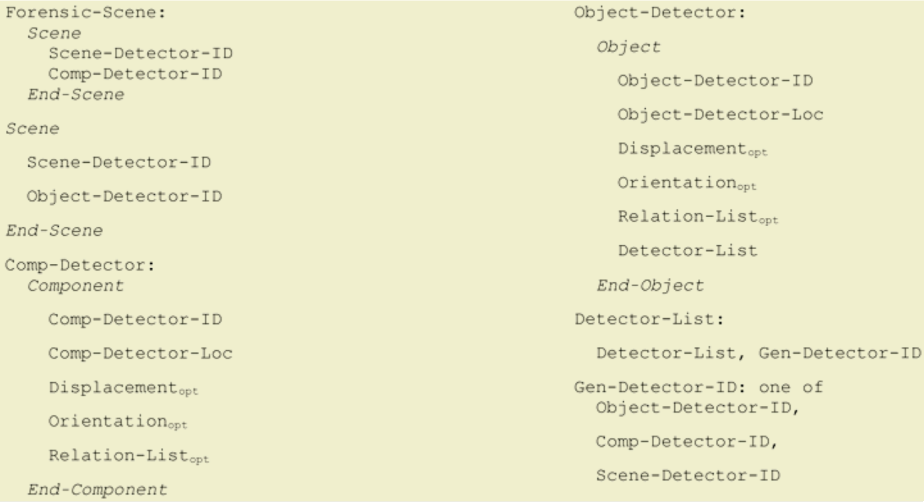


Fig. 2. Detailed diagram of the query usage process

4.1 Performance Results

One application that can benefit from our image mining system is to detect and filter out improper images such as those of partially clad people. We use this application as a case study to test the performance of this system. We use a training set of 214 images consisting of 104 positive images of partially-clad people, and 110 images of negative images of landscapes, textures, clothed people, sport scenes, etc. The patch



```
Forensic-Scene:
  Scene
    Scene-Detector-ID
    Comp-Detector-ID
  End-Scene

Scene

  Scene-Detector-ID

  Object-Detector-ID

End-Scene

Comp-Detector:
  Component

    Comp-Detector-ID

    Comp-Detector-Loc

    Displacement_opt

    Orientation_opt

    Relation-List_opt

  End-Component

Object-Detector:

  Object

    Object-Detector-ID

    Object-Detector-Loc

    Displacement_opt

    Orientation_opt

    Relation-List_opt

    Detector-List

  End-Object

Detector-List:

  Detector-List, Gen-Detector-ID

Gen-Detector-ID: one of
  Object-Detector-ID,

  Comp-Detector-ID,

  Scene-Detector-ID
```

Fig. 3. Portion of grammar developed to specify structured image queries

detectors firstly detect face, waist and pelvis; then combine these components into a hierarchy to detect partially-clad people. Fig. 4 shows a positive image with detected image patches. Each feature vector is composed of high edge coefficients defining the outline of body parts and regions of continuous tones (e.g. bare skin, texture, colour). We perform three experiments using different colour spaces and varying the use of texture homogeneity values. The first test uses HSV space, maximum value of wavelet coefficients in Hue and Value as edge coefficients, and the variance of Hue and Saturation for homogeneous regions. 92% true positive and 74% true negative detection rates are obtained. The second test uses YCbCr space, maximum values of Cb and Cr, and the variance of Cb and Cr. 79% true positive and 95% true negative detection rates are obtained. The third test is similar to the second test except that texture homogeneity values are included as features instead of the variances. The detection rates are the same as in the second test.

4.2 Discussion

From these results, we have found that HSV is more useful for finding positive images, while YCbCr is more discriminating but at a reduced rate of positive detection. The texture homogeneity is not a discriminating feature for this application. Interestingly, we observed that the skin detection using YCbCr has a similar positive rate to that of the SVM classifier. Does this imply that the rate of improvement rests with the choice of a better colour model for skin detection?

5 Grammar-Based Specification, Querying and Feedback

To facilitate the communication between forensic investigators and the system, we develop a grammar for describing objects and scenes as hierarchies of component detectors. This grammar defines the position, orientation, error bound, and spatial relationship of the components. Thus, an entire scene can be described as hierarchies at varying levels of resolution, to allow fast search of regions of interest and more detailed and computationally expensive search at a finer level. Users can use this grammar for three tasks: to specify objects and scenes for training, for querying and for providing feedback to the system. Information on the position and orientation is expressed in numerical quantities, while relative spatial arrangement can be expressed in either absolute measurements, or precise terms (e.g. north, south, east, west), or fuzzy terms (e.g. up, down, above, below). These hierarchies which can be represented in an n-ary tree data structure are encapsulated into a file grammar to support storage and manipulation for future use (see Fig. 4).

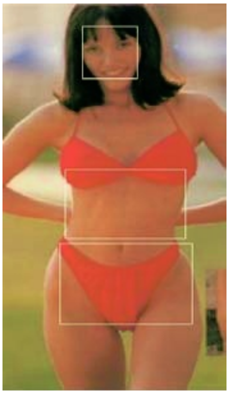


Fig. 4. An example of a positive image

This grammar is extensible to include non-spatial relationships and dynamic scenes. Non-spatial relationships would allow users to specify special characteristics of image evidence based on their previous experience. For example, the co-occurrence of bare skin and pixellated image regions might heighten the chance that the image is pornographic; the co-occurrence of weapons and important buildings might indicate a breach of security. Dynamic scenes occur in motion videos when objects may appear or disappear, or the attributes and relationships between objects may change. These changes can be implemented by appropriate operations on the n-ary tree (insertion, deletion, modification of attributes in the node contents by traversing the tree). Standard transformations (scale, translate, rotate, shear) and linguistic modifications of spatial relationships may be treated as changes in object attributes. To track an object that may be occluded from time to time, a visibility flag is used.

6 Conclusion and Future Work

We have presented a forensic image mining system which is modeled closely to the way forensic investigators work. It provides the facility for training the system to detect the image evidence required, as well as for correcting inaccurate search results or fine-tuning the search further. The communication between users and the system is facilitated by an adaptive grammar. To date, the prototype system consisting of the component-based detection engine and the grammar has been implemented and evaluated for detection of images containing partially clad humans and in other applications with very promising results [6, 7].

The system architecture is flexible in the sense that other types of classifiers (e.g. Naïve Bayes, C4.5 or neural networks) can be used instead of the SVM if they are more suited to the classification of specific types of data. Furthermore, different classifiers may be used for different parts of the system. The grammar is generic and extensible to allow more sophisticated query to be generated if required.

Bayesian networks (BN) provide a compact and efficient means to represent joint distributions over a large number of random variables and allows effective inference from observations (e.g. [18]). Hence, they can be used to understand and learn probabilistic and causal relationships through updating beliefs based on evidence provided. The need for dealing with uncertainties that are inherent in DIF has motivated the use of Bayesian networks. These uncertainties occur in image characteristics, object description, co-occurrence of objects and human semantic interpretation of image content and its relevance to forensic purposes. Our ongoing work includes the implementation of the Bayesian networks for relevance feedbacks and more extensive tests with other examples of image forensic work. It is also envisaged that subjective testing will be performed with input from forensic experts.

References

1. CSI/FBI, 2003 Computer Crime and Security Survey. Computer Security Institute, San Fransisco, USA, 2003.
2. Open Systems, www.opensystems.com.au, visited on 15 Oct. 2003.
3. W. Niblack, X. Zhu, J. Hafner et al, "Updates to the QBIC system", Storage and Retrieval for Image and Video Databases, 1997, vol. 3312, pp. 150–161.
4. G. Mohay, A. Anderson, B. Collie, O. de Vel and R. McKemmish., Computer and Intrusion Forensics, Artech House Publishers, 2003.
5. H. Muller, W. Muller, S. Marchand-Maillet et al, "Strategies for positive and negative relevance feedback in image retrieval", Proc. International Conference on Pattern Recognition ICPR2000, vol. 1, pp. 1043–1046.
6. R. Brown, B. Pham and O. de Vel, "A grammar for the specification of forensic image mining searches", Proc. 8th Australian and New Zealand Conference on Intelligent Information Systems, Sydney, Australia, 2003.
7. Brown, R.; Pham, B., "Image Mining and Retrieval Using Hierarchical Support Vector Machines", Proc 11th International, Multimedia Modelling Conference (MMM 2005), 12-14 Jan. 2005 pp. 446 – 451.
8. Yow, K. and R. Cipolla, "Feature-based human face detection", Image and Vision Computing, 1997, vol. 15, (9), pp. 713–735.
9. Haritaoglu, D. Harwood and L. Davis, "W-4: Real-Time surveillance of people and their activities", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, vol. 22, (8), pp. 809–830.
10. Camurri, M. Ricchetti and R. Trocca, "EyesWeb-toward gesture and affect recognition in dance/music interactive systems", IEEE International Conference on Multimedia Computing and Systems, 1999., Florence, Italy.
11. H. Miyamori, and S. Iisaku, "Video annotation for content-based retrieval using human behavior analysis and domain knowledge", Proc. Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 2000.

12. Tomita, A., T. Echigo, et al., "A visual tracking system for sports video annotation in unconstrained environments" International Conference on Image Processing, Vancouver, Canada, 2000.
13. N. Oliver, B. Rosario and A. Pentland., "A Bayesian computer vision system for modeling human interactions", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, vol. 22, (8), pp. 831–843.
14. Y. Ivanov, and F. Aaron, "Recognition of visual activities and interactions by stochastic parsing", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, vol 22, (8), pp. 852–872.
15. T. Wada and T. Matsuyama, "Multi-object behavior recognition by event-driven selective attention method", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, vol. 22, (8), pp. 873–887.
16. L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition.", Proceedings of the IEEE., 1989, vol. 77, (2), pp. 257–285.
17. Mohan, C. Papageorgiou and T. Poggio (2001)., "Example-based object detection in images by components", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, vol 23, (4), pp. 349–361.
18. J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, San Mateo, 1988.

Side-Match Predictive Vector Quantization

Zhen Sun, Yue-Nan Li, and Zhe-Ming Lu

Harbin Institute of Technology, Department of Automatic Test and Control,
P.O. Box 339, 150001 Harbin, China
liyuenan@dsp.hit.edu.cn, zhemingl@yahoo.com

Abstract. Vector quantization (VQ) is widely used in low bit rate image compression. In this paper, two predictive vector quantization (PVQ) algorithms that combine the concept of side-match are proposed. By controlling the quantization distortion after encoding or searching the reconstructed vector with minimum side distortion before encoding, the two proposed algorithms can decrease the quantization distortion and computational complexity respectively. The bit rates are also reduced in both algorithms by using side-math technology. The performances of the proposed algorithms are compared with several previous VQ algorithms. Simulation results have shown the efficiency of the proposed algorithms.

1 Introduction

Vector quantization (VQ) is an efficient technology for data compression and it has been successfully used in the fields of image compression, speech coding and pattern recognition. In VQ, a group of samples are quantized as a unit. Compression is archived by storing or transmitting the codeword index rather than the codeword itself. VQ is viewed as the optimal scheme for encoding stationary sources. Predictive VQ [1], [2] (PVQ), which employs a memory model is the extension of differential pulse code modulation (DPCM) in VQ [3]. Unlike memoryless VQ that quantizes each input vector independently, PVQ explores the correlations between neighboring vectors. It can be proved that PVQ can achieve more excellent compression performance than memoryless VQ if an accurate predictor is adopted.

Side match VQ [4] (SMVQ) is a sort of low bit rate VQ algorithms. In this paper, two PVQ algorithms that combine the concepts of side-match are proposed to decrease distortion and computation complexity respectively. The rest parts of this paper are organized as follows: Section 2 reviews the basic concepts of PVQ and SMVQ. In Section 3, we describe the details of the proposed algorithms. In Section 4, we give the simulation results of the proposed algorithms. Finally, the conclusions are given in Section 4.

2 Predictive VQ and Side Match VQ

2.1 Predictive Vector Quantization

The basic idea of PVQ is using a predictor to remove the predictable redundancy and then use a VQ encoder to encode the prediction error. Let's denote the input vector by $\boldsymbol{x}_n$ and its prediction value is denoted by $\tilde{\boldsymbol{x}}_n$. $\tilde{\boldsymbol{x}}_n$ is predicted by previously quantized

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 405–410, 2005.

vectors. The residual vector is denoted by $\boldsymbol{e}_n$ and it measures the difference between $\boldsymbol{x}_n$ and $\tilde{\boldsymbol{x}}_n$. The residual vector $\boldsymbol{e}_n$ is calculated as follow.

$$\boldsymbol{e}_n = \boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n \tag{1}$$

The reconstruction error of PVQ is equal to the quantization error of $\boldsymbol{e}_n$. Each component of $\boldsymbol{e}_n$ will be close to zero if an accurate prediction is implemented. As a result, PVQ can often archive higher compression performance than memoryless VQ. The basic structures of PVQ encoder and decoder are shown in Fig.1.

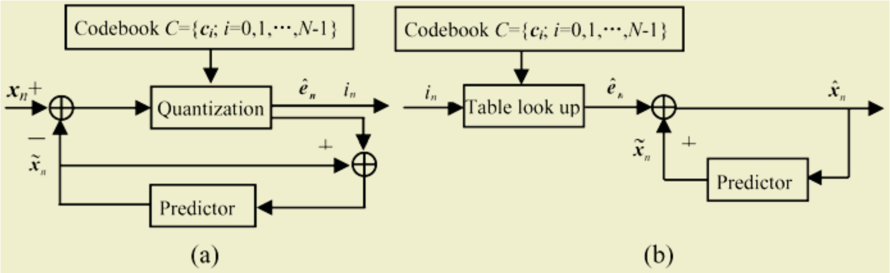


Fig. 1. Structures of PVQ encoder and decoder. (a) PVQ encoder; (b) PVQ decoder

2.2 Side Match Vector Quantization

Side match VQ (SMVQ) is a kind of finite-state VQ (FSVQ). SMVQ exploits the correlation between neighboring pixels. The side distortions between the input vector and codewords are calculated to generate the state codebook by sorting the codewords according to their side-match distortions. The codeword searching is implemented within the state codebook that is a small part of the main codebook. As a result, low bit rate can be achieved by using SMVQ. Let $\boldsymbol{y}_i$ denote the codeword in the state codebook. Then the side distortion between $\boldsymbol{x}_n$ and $\boldsymbol{y}_i$ can be calculated as follow.

$$d_{sm}(\boldsymbol{x}_n, \boldsymbol{y}_i) = \sum_{n=0}^{h-1}\left(y_{i(0,n)} - u_{(w-1,n)}\right)^2 + \sum_{n=0}^{w-1}\left(y_{i(n,0)} - l_{(n,h-1)}\right)^2 \tag{2}$$

u and l are the upper and left neighboring reconstructed vectors respectively and the vector dimension is $h \times w$. Fig.2 (a) shows the relationships between $\boldsymbol{x}_n$ and its neighboring vectors.

3 Side Match Predictive Vector Quantization

3.1 Distortion Controlled SMPVQ (DCSMPVQ)

It can be seen from the above discussions of PVQ that if an input vector is not accurately predicted the quantization error will propagate to the predictions of following vectors. Thus, it is necessary to control the quantization error at the encoder. The proposed algorithm uses side-match technology to control the distortion between the input vector and the reconstructed vector. The encoder will search the previously quantized vector that satisfies the least side-distortion criteria in the search area if the mean square error (MSE) between $\boldsymbol{x}_n$ and $\hat{\boldsymbol{x}}_n$ is larger than the threshold T. The search area is show in Fig.2 (b). As neighboring pixels are highly correlated, a conclusion can be drawn that the MSE between the two vectors is small if their side-match distortion is small. The original image is divided into 4×4 blocks to form input vectors.

The encoding process for each vector in DCSMPVQ can be described as follows.

Step 1: Encode the input vector by PVQ encoder.

Step 2: Calculate the MSE between $\boldsymbol{x}_n$ and $\hat{\boldsymbol{x}}_n$. If $d(\boldsymbol{x}_n,\hat{\boldsymbol{x}}_n) > T$, go to Step 3; otherwise, go to Step 4.

Step 3: Find the previously quantized vector $\hat{\boldsymbol{x}}_{n-k}$ in the search area that satisfying the least side-match distortion criteria to reconstruct $\boldsymbol{x}_n$. If $d(\boldsymbol{x}_n,\hat{\boldsymbol{x}}_{n-k}) < d(\boldsymbol{x}_n,\hat{\boldsymbol{x}}_n)$, then $\hat{\boldsymbol{x}}_{n-k}$ is used to reconstruct $\boldsymbol{x}_n$ and then transmit one flag bit '1' to indicate that $\boldsymbol{x}_n$ is reconstructed by the previously quantized vector; otherwise, go to Step 4.

Step 4: Set the flag bit to be '0'. The flag bit and the codeword index are sent to the decoder.

The decoding process of DCSMPVQ can be described as follows.

Step 1: Judge the flag bit of the received index. Go to Step 2 if the flag bit is '0'; otherwise, go to Step 3.

Step 2: Find the residual codeword according to the index and the residual codeword is added to the prediction value to reconstruct $\boldsymbol{x}_n$.

Step 3: Search the previously quantized vector that satisfies the least side-match distortion criteria to reconstruct $\boldsymbol{x}_n$.

3.2 Fast SMPVQ (FSMPVQ)

The main drawback of VQ is the high computational complexity. The proposed FSMPVQ reduce computation complexity by adopting the concept of side-match. The search area shown in Fig.2 (b) is also used in FSMPVQ. By searching previously quantized vector that satisfies the least side-match distortion criteria, the codeword searching process is eliminated when quantizing some of the input vectors.

Let $\boldsymbol{x}(i,j)$ denote the input vector and $\hat{\boldsymbol{x}}(m,n)$ denotes the previously quantized vector that satisfies the least side-match distortion criteria. (i,j) and (m,n) are locations of $\boldsymbol{x}(i,j)$ and $\hat{\boldsymbol{x}}(m,n)$ respectively. If $\boldsymbol{x}(i,j)$ is reconstructed by the previously quantized vector $\hat{\boldsymbol{x}}(m,n)$, the location information of $\hat{\boldsymbol{x}}(m,n)$ is added to the index as follows. First, two binary bits are added to represent $(i-m)$ that varies from {0,1,2,3}. The following bits are three binary bits to represent $(j-n)$ that varies from {-1,-2, 0, 1, 2}.

The encoding process of the proposed FSMPVQ can be described as follows.

Step 1: Search the reconstructed vector that satisfies the smallest side-math distortion criteria in search area. Go to Step 2 if $d(\boldsymbol{x}(i,j),\hat{\boldsymbol{x}}(m,n)) < T$; otherwise, go to Step 3.

Step 2: Set the flag bit to be '1'and the following bits indicate the location information of $\hat{\boldsymbol{x}}(m,n)$ as described above.

Step 3: Set the flag bit to be '0'. Encode $\boldsymbol{x}(i,j)$ by traditional PVQ encoder and then send the codeword index and the flag bit to decoder.

The decoding process of the proposed FSMPVQ can be described as follows.

Step 1: Judge the flag bit of the received index. Go to Step 2 if the flag bit is '1'; otherwise, go to Step 3.

Step 2: Find the previously quantized vector in the search area according to the location information provided by the index to reconstruct $\boldsymbol{x}(i,j)$.

Step 3: Find the residual codeword according to the index and the residual codeword is added to the prediction value to reconstruct $\boldsymbol{x}(i,j)$.

The multiple-pixels distance weighted basic prediction (MPDWBP) that proposed in [5] is adopted in this paper as the prediction scheme. The MPDWBP scheme can be expressed as follows.

$$\tilde{\boldsymbol{x}}(i,j) = \frac{i \cdot \tilde{Y}_l + j \cdot \tilde{Y}_u}{i+j} \quad (1 \le i \le 4, 1 \le j \le 4) \tag{3}$$

$$\tilde{Y}_u = 0.1X_0 + 0.2X_1 + 0.4X_2 + 0.2X_3 + 0.1X_4 \tag{4}$$

$$\tilde{Y}_l = 0.1X_0 + 0.2X_5 + 0.4X_6 + 0.2X_7 + 0.1X_8 \tag{5}$$

In the algorithm proposed in [5], memoryless VQ is used to quantize vectors that located in Row 1 and Column 1. We quantize those vectors by way of PVQ. Prediction schemes of vectors located in Row 1 and Column 1 are as follows respectively.

$$\tilde{\boldsymbol{x}}(i,j) = 0.2X_5 + 0.3X_6 + 0.3X_7 + 0.2X_8\,, \tag{6}$$

$$\tilde{\boldsymbol{x}}(i,j) = 0.2X_1 + 0.3X_2 + 0.3X_3 + 0.2X_4 \tag{7}$$

The corner vector that located in Row 1 and Column 1 is stored directly in the codebook and its reconstruction vector is the vector itself. In this way, it is not necessary to use another codebook to quantize vectors that located in Row 1 and Column 1, so the number of codewords will be reduced. The relationships between $\boldsymbol{x}(i,j)$ and its neighboring pixels are shown in the Fig.2 (c). In order to archive higher compression ratio, the Huffman coding is adopted in this paper to encode the codeword indices.

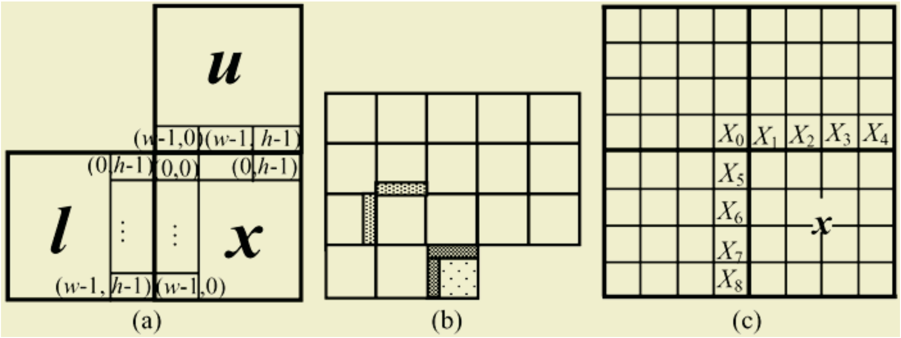


Fig. 2. SMVQ, search area and the prediction scheme. (a) Side components of input vector and codeword in SMVQ; (b) Search area. The dark block denotes the vector to be quantized and the blank ones denote previously quantized vectors; (c) The relationships between $\boldsymbol{x}$ (*input vector*) and its neighboring pixels

4 Simulation Results

Several simulations have been done to evaluate the efficiencies of the proposed algorithms. All the simulation programs are coded in the VC++ environment on the computer whose CPU frequency is 549 MHz. The Peppers and Lena images of size 512×512 with 256 gray levels are used in simulations. The Peppers image is relative smooth and the Lena image has more detail regions. The Peppers image is used to generate the residual codebook. Original images are divided into 4×4 blocks to form vectors. Codebook contains 511 residual codewords and the corner vector that located in Row 1 and Column 1 is stored directly in the codebook. The LBG algorithm is used to generate residual codebook and the initial codewords are randomly selected from the residual vectors. We select full search (FS) scheme in searching the nearest residual codeword. The performances of the proposed algorithms are compared with other VQ algorithms, including multi-stage VQ [6](MSVQ), multi-stage PVQ [7] (MSPVQ) and multiple-pixels distance weighted basic PVQ [5] (MPDWBPVQ). The simulation results of the Peppers and Lena images are listed in Table 1 and Table 2 respectively. The reconstructed Peppers and Lena images are shown in Fig.3 and Fig.4 respectively.

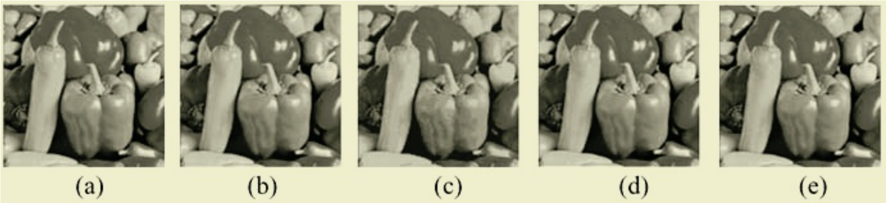


Fig. 3. Reconstructed Peppers Images. (a) DCSMPVQ *T*=400; (b) FSMPVQ *T*=300; (c) MSVQ; (d) MSPVQ; (e) MPDWBPVQ

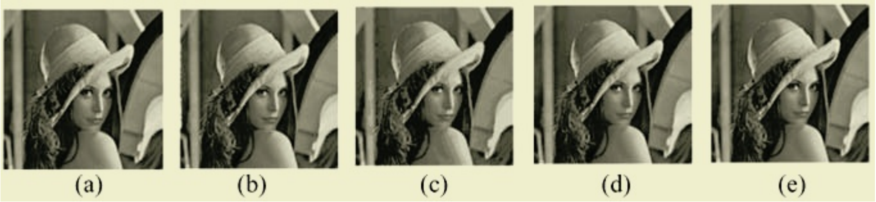


Fig. 4. Reconstructed Lena Images. (a) DCSMPVQ *T*=400; (b) FSMPVQ *T*=300; (c) MSVQ; (d) MSPVQ; (e) MPDWBPVQ

Table 1. Simulation results of Peppers image (inside the trainning set)

Algorithm		*PSNR* (dB)	*Bit rate* (bpp)	*Coding time* (s)
DCSMPVQ	(*T*=300)	32.430	0.534	2.69
	(*T*=400)	32.426	0.535	2.67
	(*T*=500)	32.423	0.536	2.66
FSMPVQ	(*T*=280)	31.992	0.481	2.10
	(*T*=300)	31.556	0.477	2.07
	(*T*=350)	31.487	0.467	1.93
MSVQ		26.442	0.562	0.72
MSPVQ		29.774	0.562	1.40
MPDWBPVQ		32.032	0.558	2.64

Table 2. Simulation results of Lena image (outside the trainning set)

Algorithm		*PSNR* (dB)	*Bit rate* (bpp)	*Coding time* (s)
DCSMPVQ	(*T*=300)	30.539	0.504	2.82
	(*T*=400)	30.534	0.506	2.81
	(*T*=500)	30.521	0.507	2.81
FSMPVQ	(*T*=280)	27.948	0.541	2.23
	(*T*=300)	27.855	0.532	2.14
	(*T*=350)	27.671	0.514	2.01
MSVQ		21.714	0.562	0.92
MSPVQ		28.523	0.562	2.11
MPDWBPVQ		29.416	0.558	2.76

5 Conclusions

Two PVQ algorithms that adopt the concept of side-match are proposed in this paper to reduce the distortion and computation complexity respectively. It can be seen from the simulation results in Table 1 that the PSNR of the proposed DCSMPVQ is 0.4dB higher than that of MPDWBPVQ. In FSMPVQ, the PSNR is less than that of DCSMPVQ as the codeword searching process is not implemented when quantizing a portion of input vectors. However, the computation complexity of FSMPVQ is less than that of MDWBPVQ. In Table 2, Lena image that is outside the training set is also encoded by several kinds of VQ algorithms. The PSNRs of the reconstructed images are lower than that of Peppers image, but the PSNR of DCSMPVQ is still higher than other VQ algorithms as distortion controlled scheme is implemented. The MSVQ and MSPVQ are two fastest algorithm and the coding time of MSVQ and MSPVQ is even less than that of FSMPVQ because of their unique encoder structures, but the reconstructed image qualities are much lower than other algorithms. The images reconstructed by MSVQ in Fig.3 and Fig.4 have serious block effects. The compression performances of the proposed algorithms are dependent on the selection of *T*. Selecting a suitable *T* can make a tradeoff between reconstructed image quality and computation complexity. For both DCSMPVQ and FSMPVQ, a smaller *T* is appropriate when encoding images with more detail regions, and a larger *T* is appropriate for images with more smooth regions.

References

1. V. Cuperman, A.Gersho: Adaptive Differential Vector Coding of Speech. In Conference Record GlobeCom 82, (1982) 1092-1096.
2. T. R. Fischer, D. J. Tinnen: Quantized Control with Differential Pulse Code Modulation. 21th Conference on Decision and Control, (1982) 1222-1227.
3. S. H. Sun, Z. M. Lu: Vector Quantization Technology and Applications. Science Press, (2002) China.
4. T. Kim: Side Match and Overlap Match Vector Quantizers for Images. IEEE Trans. on Image Processing, Vol.1, No.2 (1992) 170-185.
5. B. Yang, Z. M. Lu: Neighboring Pixels Based Low Complexity Predictive Vector Quantization Algorithms for Image Coding. ATCA ELECTRONICA SINICA. Vol.31, No.5, (2003) 707-710.
6. B. Mahesh, W. A. Pearlman: Variable-rate Tree Structured Vector Quantizers. IEEE Trans. on Information Theory, Vol.41, No.4, (1995) 917-930.
7. S. A. Rizvi, N. M. Nasrabadi: Predictive Residual Vector Quantization. IEEE Trans. Image Process, Vol.4, No.11 (1995) 1482-1495.

Improved Image Coding with Classified VQ and Side-Match VQ

Hsiang-Cheh Huang[1,2], Kang K. Yen[3], Yu-Hsiu Huang[4],
Jeng-Shyang Pan[5], and Kuang-Chih Huang[4]

[1] National Kaohsiung Marine University, Kaohsiung, Taiwan, ROC
[2] National University of Kaohsiung, Kaohsiung, Taiwan, ROC
huang.hc@gmail.com
[3] Florida International University, Miami, FL 33174, USA
Kang.Yen@fiu.edu
[4] Cheng-Shiu University, Kaohsiung 840, Taiwan, ROC
[5] National Kaohsiung University of Applied Sciences, Kaohsiung 807, Taiwan, ROC
jspan@cc.kuas.edu.tw

Abstract. A new scheme for vector quantization (VQ) is proposed in this paper. We employ side-match and classified criteria for designing VQ codebooks to combat blocking effects induced from high compression rates, and then we use the proposed algorithm to improve the reconstructed image quality. Simulation results demonstrate the better image quality to compare with that produced from conventional schemes objectively and subjectively, at the cost of reasonable encoding complexity.

1 Introduction

People easily retrieve digital images through the Internet, multimedia databases, or wireless networks nowadays. Due to the myriad amounts of image file sizes, compression is inevitable in storing and transmitting images. Unlike data compression to reduce redundancy and to pursue lossless compression, most image compression schemes belong to the lossy compression category to search for the reduction in entropy by discarding the less important portions, which are less likely to be perceived by human eyes. The reconstructed image quality is supposed to be under the just noticeable distortion (JND) region [1]. Therefore, the reconstructed image is an approximation of its original counterpart with a much lower file size. In addition, lossy compression of images often employ the human visual system (HVS) to make subjective tests and examine the effectiveness and feasibility of the relating algorithms.

Vector quantization (VQ) [2, 3], like transform coding or subband coding, is one of the useful compression schemes for images. In this paper, we propose an improved version of VQ algorithm, called combination of classified and side-match vector quantization, to look after the subjective and objective quality of the compressed image, and the encoding time of the algorithm.

This paper is organized as follows. In Section 2 and describe the fundamental concepts of VQ, and the basic ideas of classified VQ (CVQ) and side-match

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 411–417, 2005.

VQ (SMVQ). In Section 3 we address the redesign of VQ codebook algorithm, which is a combination of CVQ and SMVQ. Simulation results are presented in Section 4. Finally we summarize and conclude our major findings in Section 5.

2 Fundamental Concepts

We give some fundamental concepts in VQ in Sec. 2.1. Then, we introduce two variations in VQ, namely, classified VQ in Sec. 2.2, and side-match VQ in Sec. 2.3, which are served as the building blocks of the proposed algorithm.

2.1 Basics of VQ

Vector quantization, as an extension to scalar quantization, works on vectors of raw data. A vector can be a small block of image data, for example, the grey-level values of a 4 × 4 pixel image block forms a 16-dimensional vector. All the blocks are encoded independently; however, there are many inter-block correlations among neighboring blocs. How to use information of adjacent blocks to further reduce the bitrate or filesize is the main concern in Secs. 2.2 and 2.3. Fig. 1 gives a block diagram for vector quantization compression.

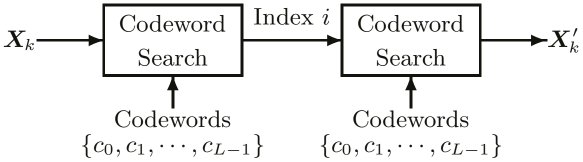

X_k
Codeword Search
Index i
Codeword Search
X'_k
Codewords $\{c_0, c_1, \cdots, c_{L-1}\}$
Codewords $\{c_0, c_1, \cdots, c_{L-1}\}$

Fig. 1. A block diagram for vector quantization. All the L codewords form a codebook with length L. $\boldsymbol{X}_k$ means the vector, and $\boldsymbol{X}'_k$ is the reconstructed vector

The original image $\boldsymbol{X}$ is composed of the combination of all the input vectors, $\boldsymbol{X}_k$, ∀ k. In the sender end, the codeword search process looks for a "nearest codeword," c_i, from the codebook for the given input vector $\boldsymbol{X}_k$. The codebook $\boldsymbol{C}$ with length L is composed of L elements $\boldsymbol{C} = \{c_0, c_1, \cdots, c_{L-1}\}$. Euclidean distance is employed in the search process to measure the distance between the two vectors, the codeword c_i and the input vector $\boldsymbol{X}_k$, as indicated in Eq. (1),

$$i = \arg\min_j D\left(\boldsymbol{X}_k, c_j\right), \quad j = 0, 1, 2, \cdots, L-1, \tag{1}$$

where $D(\bullet, \bullet)$ denotes the Euclidean distance.

The index of selected codeword i is then transmitted to the receiver end. With the same codebook $\boldsymbol{C}$, the decompression process can easily reconstruct vector $\boldsymbol{X}'_k$ by simple table look-up, as depicted in Fig. 1. All the reconstructed vectors, $\boldsymbol{X}'_k$, ∀ k, make up the reconstructed image, $\boldsymbol{X}'$.

The codebook plays an essential role in VQ. The codebook size is a tradeoff between the reconstructed image quality and the compression rate. The codewords in the codebook decide the resultant compression distortion. A dedicated procedure is required for the generation of appropriate codebook. Among other alternatives, LBG algorithm [3] is widely used in various applications.

2.2 Backgrounds for Classified VQ (CVQ)

Human eyes are sensitive to local contrasts between bright and dark luminance [4]. In CVQ, it employs multiple codebooks, and each each one has its own characteristics, for example, horizontal edges, vertical edges, or diagonal edges. By using smaller codebooks, each is designed for some specific characterics of each block, CVQ leads to less complexity and less computation time to compare to conventional VQ schemes such as those in Fig. 2.

Fig. 2 shows the 28 binary edge classes [4]. For every 4 × 4 block in an image, the threshold T_0 is determined by

$$T_0 = \sum_{i=1}^{4} \sum_{j=1}^{4} \frac{w(i,j)}{W} g(i,j) \tag{2}$$

where

$$W = \sum_{i=1}^{4} \sum_{j=1}^{4} w(i,j) \tag{3}$$

and $w(i,j)$ is the weighting factor associated with the grey value $g(i,j)$. By using thresholding such that

$$g(i,j) = \begin{cases} 1, & \text{if } g(i,j) \geq T_0, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

it becomes a data reduction process which separates the block into two classes of objects. Each shade block is based on the weight $w(i,j)$. For a preset threshold T_s, if any $w(i,j) \geq T_s$, it is considered non-shade; otherwise, it is considered a shade block. Once the block is considered non-shade, it is transformed into its binary representation, and it is classified into one of the 28 classes in Fig. 2. With CVQ, each class has its own codebook, which produces effective reconstruction with better image quality. More details about CVQ can be found from [4].

2.3 Backgrounds for Side-Match VQ (SMVQ)

Side-match VQ (SMVQ) is a well-known class of finite state vector quantization (FSVQ), which is adopted for low-bit rate image coding. It exploits the spatial correlation between the neighboring blocks to select several codewords that are very close to the encoding block from the master codebook.

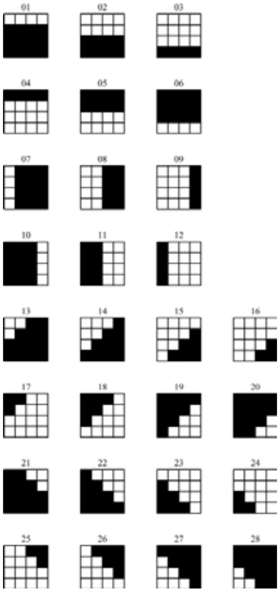


Fig. 2. Binary edge classes

SMVQ uses the master codebook $\boldsymbol{C}$ to encode blocks in the first column and the first row in advance, and the rest of blocks are recovered by side match. For a codeword $y \in \boldsymbol{C}$, we calculate the side-match distortion (smd) between the current $k \times k$ block, and its upper-adjacent block u and left-adjacent one l, by

$$\mathrm{smd}(y) = \sum_{i=0}^{k-1} \left(u_{(k-1,i)} - y_{(0,i)}\right)^2 + \sum_{i=0}^{k-1} \left(l_{(i,k-1)} - y_{(i,0)}\right)^2 . \tag{5}$$

Then, it sorts the master codebook $\boldsymbol{C}$ according to side-match distortions of all codewords therein, and selects the first N_s codewords to form the state codebook, $\boldsymbol{SC}$, with the smallest side-match distortions. With SMVQ, it can obtain a better compression capability.

3 Proposed Algorithm with CVQ and SMVQ

For VQ compression, every block in an image is independently encoded with VQ. Due to the inter-block correlations and spatial redundancies, with high compression ratios, the raggedness within one block and the blocking effect introduced among adjacent block would become perceptually noticeable that need to be conquered. Hence, we combine both the advantages of CVQ and SMVQ to propose our algorithm demonstrated in Figs. 3 and 4.

In Fig. 3, by pre-setting the thresholds for CVQ, TH_c, and SMVQ, TH_s, we can obtain the state codebook and block classifier. In Fig. 4, we encode each

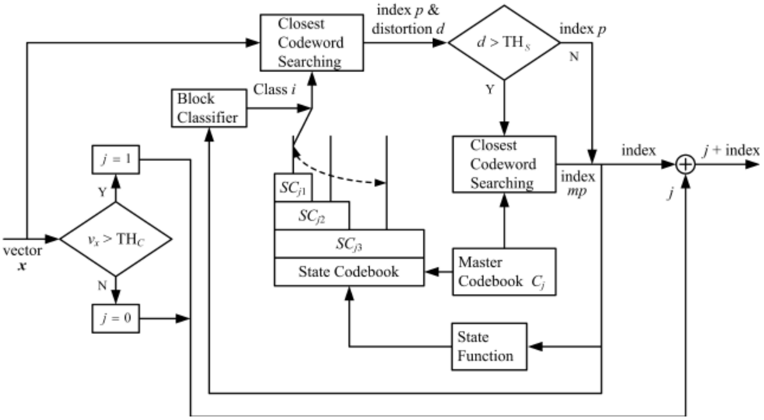


Fig. 3. The flow chart of the proposed algorithm

X	O	X	O	⋯ ⋯	X	O
O	X	O	X	⋯ ⋯	O	X
X	O	X	O	⋯ ⋯	X	O
O	X	O	X	⋯ ⋯	O	X
⋮	⋮	⋮	⋮	⋱	⋮	⋮
X	O	X	O	⋯ ⋯	X	O
O	X	O	X	⋯ ⋯	O	X

Fig. 4. Encoding orders with our algorithm. Each square means a 4 × 4 block. Blocks with X means encoding by CVQ, while those with O means encoding by SMVQ

block with CVQ and SMVQ alternatively. By doing so, we hope to obtain a tradeoff among compression ratio, computation complexity, and reconstructed image quality.

4 Simulation Results

In our simulation, we take the test image, `Lena`, with size 512 × 512, as the original source. It is divided into 4 × 4 block for VQ compression. By following Sec. 3, parameters employed are stated as follows. In SMVQ, we set the state codebook size $N_s = 16$, $\mathrm{TH}_s = 500$, and the size of the master codebook is 29 × (the number of non-shade blocks). In CVQ, $T_s = 9$, and the non-shade codebook size is 29×6 in Sec. 3.

Three objective metrics are employed to examine the effectiveness and advantage of the proposed algorithm. These are:

1. *Reconstructed image quality.* Reconstructed image quality is measured by Peak Signal-to-Noise Ratio (PSNR) between $\boldsymbol{X'}$ and $\boldsymbol{X}$. The objective quality of watermarked image is better if we have the larger PSNR value.
2. *Compression ratio.* We calculate the compressed bit rate, represented by 'bit per pixel,' in our simulations. Because the original image is 8-bit/pixel, we can easily calculate the compression ratio by dividing the two amounts.
3. *Consumed computation time.* We record the time consumed in compressing the images with different algorithm. It indicates that the longer time consumed, the more complexity introduced.

Simulation results with our scheme are presented in Fig. 5 for subjective comparisons, and in Table 1 for objective evaluations.

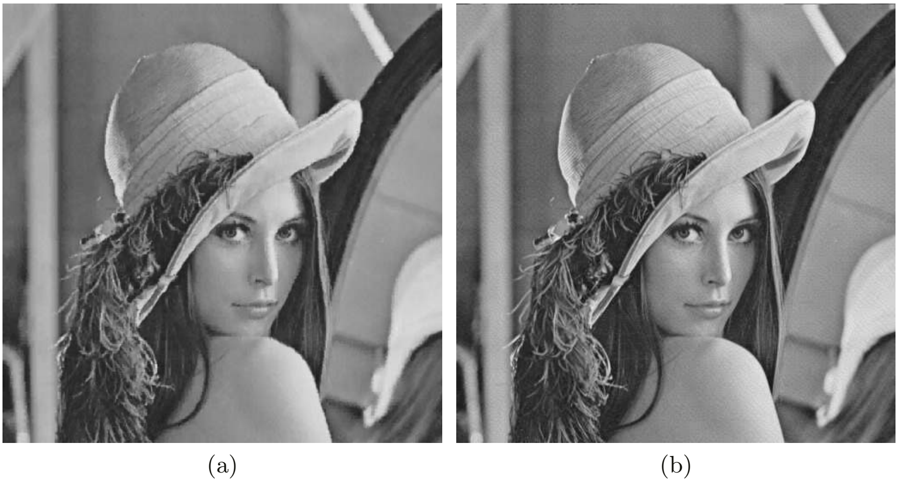


Fig. 5. Comparisons of subjective image quality. (a) The original image `lena` with size 512 × 512. (b) The reconstruction generated from the proposed algorithm in Sec. 3

Table 1. Comparisons of the existing scheme and the schemes proposed in this paper

Schemes	Proposed	CSMVQ [5]	VCSMVQ [6]
PSNR (dB)	33.74	32.07	31.49
bit/pixel	0.62	0.61	0.63
Compression ratio	12.90	13.11	12.70
Computation time (sec)	23.9	0.7	77.59

After evaluating the metrics above, the reconstructed image quality with our proposed algorithm outperform those with existing algorithm in [5] and [6]. The compression ratios, represented by bit per pixel, are very similar with the three algorithms. The computation time with [5] is the smallest due to its simplicity in encoding, while the one with [6] is the largest due to its complexity. The computation time with our algorithm is in the middle of the two.

To sum up, we conclude that the image quality with our algorithm outperforms those from other algorithm, with somewhat longer computation time consumed to compare with the one with the least computation time. This means the effectiveness of the proposed algorithm.

5 Conclusion

We proposed an improved scheme for VQ compression in this paper. By combining the advantages of CVQ (better reconstructed quality) and SMVQ (better compression capability), we proposed an improved VQ compression algorithm with better subjective image quality. Also, simulation results presented that under similar compression ratios, the image quality with our algorithm outperforms those from existing algorithms, with a reasonable computation time consumed. Further studies, such as error resilient transmission of VQ compressed images in [7], will be exploited in the future.

Acknowledgements

This work is partially funded by Dept. of Defense (USA) under award number H98230-04-C-0460, Dept. of Transportation (USA) under project number FL-29-7102-00 and National Science Council (Taiwan, ROC) under grant number NSC 93-2213-E-151-016. The authors would also like to thank Mr. Chih-Wei Hsu and Mr. Chao-Chin Chang for their helps in programming.

References

1. N.S. Jayant, J.D. Johnston, and R.J. Safranek, “Signal compression based on models of human perception,” Proc. IEEE, vol. 81, no. 10, pp. 1385-1422, Oct. 1993.
2. A. Gersho and R.M. Gray, Vector Quantization and Signal Compression. Kluwer Academic Publishers: Boston, MA, 1992.
3. Y. Linde, A. Buzo, and R.M. Gray, “An algorithm for vector quantizer design,” IEEE Trans. Commun., vol. 28, no. 1, pp. 84-95, Jan. 1980.
4. M.K. Quweider and E. Salari, “Efficient classification and codebook design for CVQ,” IEE Proc. Vision, Image and Signal Process., vol. 143, no. 6, pp. 344–352, Dec. 1996.
5. R.F. Chang and W.T. Chen, “Image coding using variable-rate side-match finite-state vector quantization,” IEEE Trans. Image Process., vol. 2, no. 1, pp. 104–108, 1993.
6. Z.M. Lu, J.S. Pan and S.H. Sun, “Image coding based on classified side-match vector quantization,” IEICE Trans. Inf. & Syst., vol. E83-D, no. 4, pp. 2189–2192, Dec. 2000.
7. N. Farvardin, “A study of vector quantization for noisy channels,” IEEE Trans. Inform. Theory, vol. 36, no. 4, pp. 799–809, 1990.

Fast Multiple Reference Frame Motion Estimation for H.264 Based on Qualified Frame Selection Scheme

Tien-Ying Kuo and Huang-Bin Chen

Dept. of Electrical Engineering, National Taipei University of Technology, Taiwan, R.O.C.
tykuo@ee.ntut.edu.tw, hbchen@image.ee.ntut.edu.tw

Abstract. Multiple reference frame motion compensation has been adopted by the emerging video coding standard H.264/AVC. But the motion estimation at the encoder over multiple reference frames to find the best inter coding is slow and computationally involved. Thus, a fast algorithm for reference frame selection and motion estimation is proposed to reduce the complexity. The proposed method, Fast Multiple Reference Frame Motion Estimation (FMRFME) selects the suitable reference frames according to the initial motion search results of 8x8 block size, and only the selected frames should be further tested in variable block size motion estimation. The experimental results show that the proposed method reduces a considerable amount of complexity of multiple reference frame motion estimation while keeping the same R-D performance as full search.

1 Introduction

The latest video coding standard, H.264/MPEG-4 AVC [1], introduces several new coding tools to improve the rate-distortion performance of the past coding standards. The enhancement for inter-frame coding in H.264 includes quarter-pixel accuracy, variable block size partitioning, and multiple reference frame motion compensation, etc. Multiple reference frames motion compensation allows the encoder to predict a picture using many stored pictures which had been coded previously. It achieves better performance in the cases of repetitive motions, uncovered background, non-integer pixel displacement, and lighting changes, etc. However, the computational complexity of motion estimation process increases dramatically in comparison with that of one reference frame, which is adopted in previous standards. Variable block sized motion compensation in H.264 supports seven different block sizes (16x16, 16x8, 8x16, 8x8, 8x4, 4x8, 4x4), which are defined as mode 1 to mode 7 respectively. Variable block size motion compensation can represent the motion characteristic in a macroblock more accurately, and therefore reduce the prediction error. However, a macroblock could have possible 259 combinations of partition under such seven modes, and this will radically increase the motion search complexity, especially, combined with multiple reference motion estimation. The main objective of this paper is to investigate this problem and propose a fast multiple reference frame motion estimation to reduce such huge complexity.

In literature, there were several works proposed for complexity reduction of multiple reference frame motion estimation. Duanmu [2] proposed the continuous tracking

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 418–424, 2005.

algorithm by exploiting the correlations between the frame-dependent motion vectors in the time domain, and Su [3] re-used the motion information of each frame to its previous frame for obtaining the long-term motion vectors. These two methods can provide better initial search candidates to speed-up multiple reference frame motion estimation. However, the motion tracking across frames may not always sustain, and every reference frame still should be searched. In [4], Chang skipped partially the motion estimation by excluding some reference frames based on the information of the sub-pixel motion displacement of each block. However, Chang's method didn't solve the variable block size problem in multiple frames. Huang [5] decided whether the encoder should search most recent reference frame or more remote frames according to many pre-set thresholds using the results of motion estimation in previous frame and intra prediction. However, the thresholds are sequence dependent. In this paper, an efficient multiple reference frame selection method, Fast Multiple Reference Frame Motion Estimation (FMRFME), was proposed. The proposed method selects the suitable reference frames according to the initial search results of 8x8 block size. Therefore, only the selected qualified frames should be further tested in motion estimation. We will provide the detailed description of the proposed method in Section 2. Experimental results and conclusions are given in Sections 3 and 4 respectively.

2 Proposed Method

The flowchart of the proposed method is illustrated in Fig. 1. At first, a 16x16 MB is disjoined into four blocks of size 8x8 (i.e., mode 4) to perform diamond search [6] on the immediate previous frame, followed by the sub-pixel motion refinement to obtain the four motion vectors, MVs^{t-n} (the value of frame distance n here is one). If all MVs^{t-1} are zero, the macroblock has high probability to be static content, and our encoder will test modes 1, 2, 3, and 4 only on previous frame for inter prediction. On the other hand, if one of MVs^{t-1} is non-zero, another strategy, continuous tracking diamond search (CTDS), will be used to search on all the rest reference fames with mode 4, i.e., four 8x8 blocks. Based on the CTDS search results, we can exclude the unqualified reference frames and modes by the value of MVs^{t-n} of 8x8 blocks. Only the frames referred by MVs^{t-n} will be performed on variable block size mode test, and some of the test modes may be skipped depending on the variance of MVs^{t-n}. If the variance of four MVs^{t-n} is zero, it represents that the contents in a MB have uniform motion characteristics, and only larger block size modes, modes 1, 2, 3, and 4, on the corresponding frame are performed. Finally, a fast motion estimation approach is always proposed by properly setting the initial search position and search range for each selected reference frame with valid modes.

2.1 Continuous Tracking Diamond Search (CTDS)

The purpose of CTDS is to find out the best motion vectors of mode 4 (block size 8x8) from current frame to the reference frames. The CTDS result is passed to our multiple reference frame selection scheme discussed in the next sub-section. Due to a

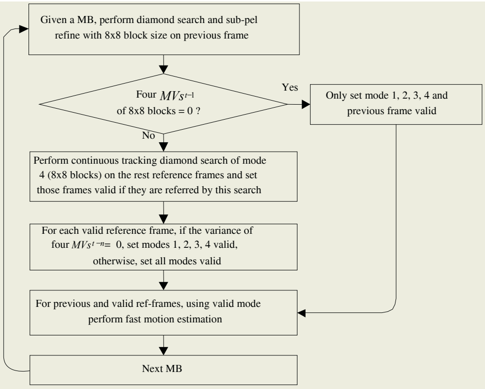

Given a MB, perform diamond search and sub-pel refine with 8x8 block size on previous frame
Four MVs^{t-1} of 8x8 blocks = 0 ?
Yes
Only set mode 1, 2, 3, 4 and previous frame valid
No
Perform continuous tracking diamond search of mode 4 (8x8 blocks) on the rest reference frames and set those frames valid if they are referred by this search
For each valid reference frame, if the variance of four MVs^{t-n}= 0, set modes 1, 2, 3, 4 valid, otherwise, set all modes valid
For previous and valid ref-frames, using valid mode perform fast motion estimation
Next MB

Fig. 1. The flowchart of the proposed method

great deal of correlation between video frames, the motion vectors to two successive reference frames could be very similar. As shown in Fig. 2, if we already obtain the motion vector to time *t-1*, we can use it as the initial search point to search at time *t-2*, and track it accordingly in this fashion until the farthest reference frame at time *t-4* is reached. In this way, the tracking can provide a very good initial search position to find out the motion vectors of successive reference frames rapidly. In Duanmu's continuous tracking algorithm [2], it only searched a small area around the tracking motion vectors and this method might be failed if the target vector has larger displacement away from the tracking. In our proposed algorithm, the diamond search strategy [6] is used to refine the tracking to avoid continuous large motion situation, and still keep the search complexity as low as possible. The initial search position of diamond search on Frame *t-n* can be represented as:

$$InitialSearchPos(x,y,t-n) = (MVx_t^{t-n+1}, MVy_t^{t-n+1}) \tag{1}$$

where MVx_t^{t-n+1} represents the x direction component of motion vectors of current frame referring to Frame *t-n+1*, and so does y direction component.

2.2 Multiple Reference Frame Selection (MRFS)

Based on the vector output of CTDS, multiple reference frame selection (MRFS) strategy can effectively drop the unqualified reference frames and prevent they from further motion search for other modes. The concept of the proposed MRFS is illustrated in Fig. 3. In the left part of Fig. 3, supposed that the motion vectors for four 8x8 blocks (mode 4) searched by CTDS are located on time *t-1*, *t-2*, and *t-4*. Our MRFS strategy is to drop the non-referred frames, Frame *t-3* in this example, and only perform all the other modes of motion search on the referred frame in CTDS, that is Frames *t-1*, *t-2*, and *t-4*. We will expect MRFS will speed up the motion search if unqualified frames are dropped under the condition that the right reference frames are kept. Supposed we perform full search on all the reference frames and the outcome is not related to Frame *t-3*, as shown on the right side of Fig. 3. Then, our MRFS saves complexity by dropping a right frame, and thus it has 100% hit rate. We test the average CTDS hit rate in Table 1. In Table 1, $P(E_{8x8})$ denotes the hit rate of MRFS to the full search, for the event E_{8x8} with four 8x8 motion vectors for a macrobock used. $P(E_{4x4})$ denotes the same meaning while using sixteen 4x4 blocks (i.e., mode 7) in a marcoblock. Table 1 shows that the hit rate is high around 92%-97% if we use four 8x8 blocks. Thus we design the 8x8 block motion search in our CTDS in terms of both hit rate and speed.

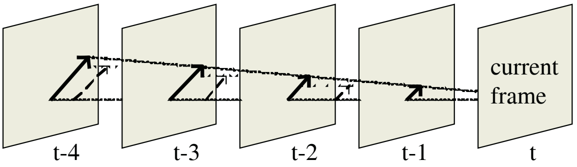


Fig. 2. Motion vector correlation in successive frame

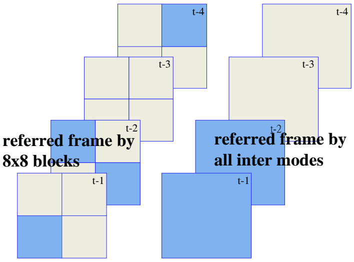

t-4
t-3
t-2
t-1
referred frame by 8x8 blocks
t-4
t-3
t-2
t-1
referred frame by all inter modes

Fig. 3. Illustration of reference frame correlation

Table 1. Hit rate of the proposed multiple reference selection strategy to the full search

Sequence	$P(E_{8x8})$	$P(E_{4x4})$	#AVG-REF / MB
Suzie	97.05%	96.40%	1.395
Stefan	92.08%	90.40%	1.758
Foreman	93.87%	92.67%	1.737
Carphone	94.87%	93.90%	1.747
Mobile	93.97%	89.51%	2.291
Average	**94.37%**	**92.58%**	**1.786**

Note that, even with the high hit rate, we expect that MRFS can drop as many frames as they can drop. The last column in Table 1 indicates the average number of qualified reference frames kept per MB after MRFS decision. As shown in this column, the encoder can save a lot of computation complexity from 5 reference frames down to 1.786 reference frames in average.

2.3 Fast Motion Estimation on Variable Block Size Modes

The best variable block size mode should be checked on those qualified reference frames selected by MRFS. If the variance of MVs^{t-n} given by CTDS is zero, the block content is uniform, then only the larger block partitions (macroblock partition), i.e., the mode 1 to 3 should be further checked. However, if the variance of MVs^{t-n} is not zero, we will check all modes, including both the macroblock and sub-macrolock partitions. In order to efficiently perform motion estimation for all partitions, we use the motion information obtained in CTDS process. For 16x16, 16x8, and 8x16 partitions, down-to-up prediction is used. The initial search position for each selected frame of different block types is set as the mean of MVs^{t-n} inside the blocks, and the search range is the max difference value among MVs^{t-n}. For sub-macroblock partition, 8x8, 8x4, 4x8, and 4x4 blocks, we use up-to-down prediction. The initial search position for each selected frame of different block types comes from the MVs^{t-n} value of corresponding 8x8 blocks, and the search range is the half value of $\max\left|MVs^{t-n}\right|$. In summarize, for each valid reference, we design efficiency initial search position and reduce search range for different block modes.

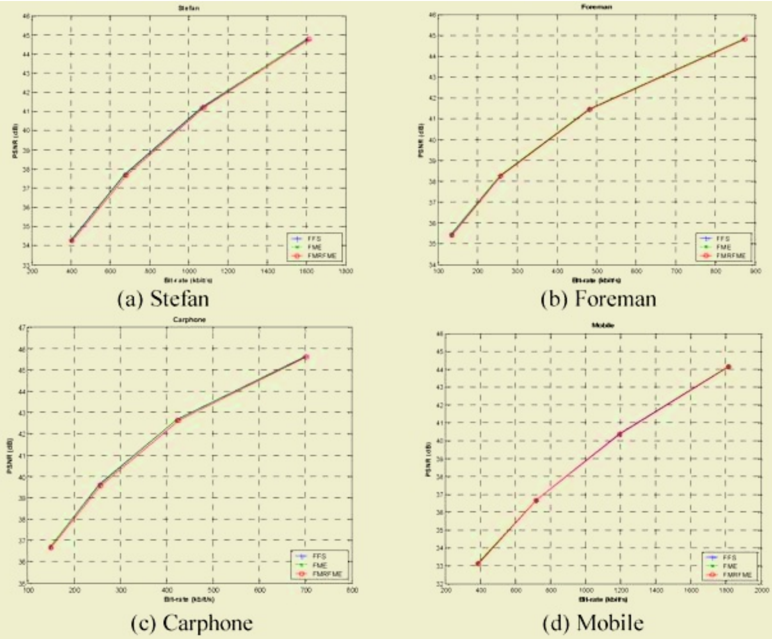
3 Experimental Results

The proposed method is implemented on the H.264 reference software JM9.2 [7]. Five QCIF video sequences are used in the experiment. For all the simulations, we use the baseline profile and turn on the RDO with four QP value including 28, 24, 20, and 16. Here the number of multiple reference frames is set to 5. And, the experiment is run on the Intel® Pentium 4 2.7GHz with 512 MB ram.

First, we compare the rate-distortion(R-D) performance of the proposed method, FMRFME, to those of the fast full search FFS and the fast search FME methods in JM9.2. The R-D curves generated by different methods with four sequences are plotted in Fig. 4. In Fig. 4, both fast methods, our proposed FMRFME and JM's FME, can keep the similar R-D performance to FFS. However, the complexity of these three methods are different. The complexity evaluation of FME and FMRFME is shown in Table 2. The encoding time is used for evaluating the speed-up factor. We measure both the total encoding time and the encoding time spent on motion estimation. As shown in Table 2, the average encoding speed-up of FMRFME with RDO on is about 2.69 times over FFS, and the motion estimation time speed-up is more about 20.47 times in average. It also shows that the proposed FMRFME method reduces considerable complexity percentage of motion estimation in encoder end from 54.63% down to 8.71% in comparison with FFS. Note that the speed-up factor of our proposed FMRFME is better than FME in all cases.

Table 2. Speed-up factors of FME and FMRFME to FFS

	Motion Estimation Time / Encoding Time			**Encoding Time Speed-Up**		**Motion Estimation Time Speed-Up**	
Sequence(frames)	FFS	FME	FMRFME	FME	FMRFME	FME	FMRFME
Suzie(150)	56.43%	14.71%	6.28%	2.26	3.11	8.68	27.99
Stefan(300)	53.26%	17.97%	15.41%	1.97	2.38	5.85	8.27
Foreman(400)	55.21%	16.68%	8.42%	2.11	2.74	6.98	17.96
Carphone(350)	55.63%	15.00%	9.30%	2.18	2.73	8.10	16.35
Mobile(300)	52.60%	15.79%	4.12%	1.95	2.49	6.51	31.8
Average	54.63%	16.03%	8.71%	2.09	2.69	7.22	20.47



(a) Stefan

(b) Foreman

(c) Carphone

(d) Mobile

Fig. 4. Rate-Distortion curves comparison among FFS, FME and FMRFME

4 Conclusion

In this work, a fast multiple reference frame motion estimation, FMRFME, is proposed to alleviate the huge complexity resulting from the multiple reference frames and block partitions. The proposed FMRFME does not search on all reference frames or all types of block partitions. The unqualified reference frames are filtered out by the output of the continuous tracking diamond search, while valid modes are also chosen based on its statistic output. The experiment demonstrates that the proposed method significantly reduces the complexity of motion estimation at encoder end for H.264 with a average factor of 1/20 over full search FFS while keeping almost the same R-D performance as that of FFS at different bit rates and different motion sequences. Hence, we conclude that the proposed FMRFME is more effective and suitable for real time applications.

Acknowledgements

This work was supported by the Nation Science Council (NSC 93-2213-E-027-041-, NSC 93-2219-E-002-031-).

References

1. Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, "Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC)", ITU-T | ISO/IEC, 2003.

2. Duanmu, C.; Ahmad, M.O.; Swamy, M.N.S.: A continuous tracking algorithm for long-term memory motion estimation [video coding], Circuits and Systems, 2003. ISCAS'03. Proceedings of the 2003 International Symposium on Volume 2, 25-28 May 2003 Page(s):II-356 - II-359 vol.2.
3. Yeping Su; Ming-Ting Sun: Fast multiple reference frame motion estimation for H.264, Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on Volume 1, 27-30 June 2004 Page(s):695 - 698 Vol.1.
4. Chang, A.; Au, O.C.; Yeung, Y.M.: A novel approach to fast multi-frame selection for H.264 video coding, Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on Volume 3, 6-10 April 2003 Page(s):III - 413-16 vol.3.
5. Yu-Wen Huang; Bing-Yu Hsieh; Tu-Chih Wang; Shao-Yi Chient; Shyh-Yih Ma; Chun-Fu Shen; Liang-Gee Chen: Analysis and reduction of reference frames for motion estimation in MPEG-4 AVC/JVT/H.264, Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on Volume 3, 6-10 April 2003 Page(s):III - 145-8 vol.3.
6. Jo Yew Tham, Ranganath, S., Ranganath, M., and Kassim, A.A.: A novel unrestricted center-biased diamond search algorithm for block motion estimation", Circuits and Systems for Video Technology, IEEE Transactions on, Volume 8, Issue 4, Aug. 1998 pp.369 – 377.
7. JM Reference Software 9.2, http://bs.hhi.de/~suehring/tml/.

Block Standstill and Homogeneity Based Fast Motion Estimation Algorithm for H.264 Video Coding

Feng Pan[1], H. Men[2], and Thinh M. Le[2]

[1] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
efpan@i2r.a-star.edu.sg
[2] National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260
{eng10921,elelmt}@nus.edu.sg

Abstract. The paper proposed a fast motion estimation algorithm based on the spatial homogeneity and temporal standstill of video sequences. It classified a frame into different regions such as standstill, stationary or dynamic region. For MBs lying in the standstill region, motion estimation is skipped; for MBs lying in the stationary region, motion estimation is applied once to an MB in the region, and the resulting motion vector will be used by other MBs in the same region. Normal motion estimation is only carried out in the dynamic region. The new algorithm can be used either by itself alone, or by combining with other existing fast motion estimation methods. Experimental results show that the new algorithm can significantly improve the time efficiency of the H.264 encoder, and is able to achieve an average reduction of 27% in encoding, with an average PSNR loss of only 0.09 dB and 0.42% bit rate increase, compared with the original H.264 reference software.

1 Introduction

The new H.264/AVC video coding standard [2] has been proved to greatly outperform MPEG-4 part 2 and H.263 standards, in terms of both PSNR and visual quality [3]. This is due to the new techniques, such as spatial prediction in intra coded blocks, integer transform, variable block size motion estimation/compensation (ME/C), multiple reference frame motion estimation/compensation, loop filter, and context adaptive binary arithmetic coding (CABAC), etc. used in the standard. Among these techniques, rate distortion optimization (RDO) is one of the essential parts of the whole encoder to achieve the much better coding performance in terms of minimizing compressed video data bits and maximizing coding quality. In RDO, the encoder tries all possible mode combinations such as different block sizes for intra prediction, inter prediction, multiple-reference frames in the case of inter modes and chooses the best one in terms of least RDO cost. This requires a lot of computational resources.

As specified in H.264, there are conceptually 7 different block sizes (16×16, 8×16, 8×16, 8×8, 8×4, 4×8 and 4×4) that can be used in, inter prediction modes. By exhaustively trying all the possible sizes, motion estimation and RDO are performed to find the best block sizes in the macroblock, resulting in heavy computational load at the encoder.

In order to reduce the temporal computational complexity, there are two interdependent levels of complexity reduction proposed in the literature: RDO and ME. At the RDO level where the decision is made based on which of the 10 search modes

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 425–432, 2005.

results in the least RDcost. Some tried to rearrange the RDO structure such that only the most likely modes are involved in the RDO computation, and the unlikely modes can be skipped [4]. At the ME level, three main approaches have been reported. The first approach attempts to reduce the number of search locations in the search area based on the assumption that matching distortion increases monotonically as the displaced candidate block moves away from the direction of the minimum distortion [7-9] The second approach proposes to reduce the number of pixels involved in the distortion calculation (while maintaining the same number of search locations) based on the assumption that object displacement is constant within a small 2-D block of pixels [10-11]. The third approach reduces the bit-depth of the involved frames and the estimation process can be done on the binary edge images. The underlying assumption of this group is that pixel illumination between frames is spatially and temporally uniform [12].

As video-conferencing becomes more and more widely accepted nowadays, it is noticed that in these typical head-shoulder video sequences, only the foreground persons are moving on top of the large portion of stationary background. Despite the zero motions in these static areas, the various motion estimation algorithms still go through all the necessary search points to find the best match, which results a inefficient use of computational resources. Thus, we propose an algorithm to first classify blocks as static and moving, and secondly apply any suitable ME technique to the moving blocks. This algorithm efficiently bypass the motion search step once a block is identified as static in motion, and assign it with either a zero-displacement motion vector; for MBs in the stationary region, the motion vector is predicated from that of the MBs. The examined MB is identified as being static by its standstill and homogeneity information. Frame differencing is adopted to decide the standstill of the MB and edge information is used to judge its homogeneity. The algorithm proposed is capable of reducing 27% of total encoding time on average with negligible PSNR lost of 0.09 and 0.42% increase in bit-rates. The rest of the paper is organized as follows. Section 2 explains how the standstill and homogeneity are determined; section 3 presents the detailed algorithm implementation. Experimental results will be presented in section 4, and conclusions will be given in section 5.

2 Standstill and Homogeneity Regions Determination

2.1 Standstill Determination

Standstill refers to the state where there is no motion among neighboring frames. For many head-shoulder video sequences, there are large portion of stationary backgrounds over the frames temporally. It is found that the motion vectors are mostly zero or same as the neighboring MBs. Thus, we can use frame difference to first judge whether the examined MB moved over time. If after subtraction, all the points within an MB have zero values, it shows that particular MB is in absolute stationary state, and it will be exempted from the motion search.

However, due to reasons like variation in background illumination and white noise, perfect zero difference is quite rare in real life, even for stationary MBs. Thus, a preset threshold Thd_S can be used to gauge whether it is a stationary MB.

2.2 Homogeneity Determination

Frame difference method can show very good results for several sequences, for example, Akiyo and News; as these sequences exhibit little noise, and illumination is rather constant. However, for sequences like Grandma, and Mother and daughter, although large stationary backgrounds exit, the frame difference method failed to segment it out. This is due to the fact that the stationary and yet uniform background is very sensitive to noise interference and variation in illumination. At the same time, it is noticed that there are many such head-shoulder video sequences in which the background tends to be homogeneous. Thus, homogeneity can be another cue to detect the background regions.

There exit many techniques for determining homogeneous regions in an image [13] [14]. A region is homogeneous if the textures in the region have very similar spatial property. The simplest method is to use statistical measures such as standard deviation, variance, skewness, and kurtosis [12]. In [13], texture is modeled using Gaussian Markov Random Field. The different textures are labeled separately using a hypothesis-and-test-based method on variable window sizes of the textures. This technique is very effective but is computationally intensive. Therefore, the ideal technique chosen should be able to detect homogeneous region effectively, while at the same time must also have low time-complexity. An effective way of determining homogeneous regions is to use the edge information, as the video object boundary usually exhibits strong edges, and we use the edge information as a supplement for detecting stationary homogeneous background. It is worth mentioning if the fast INTRA mode decision algorithm [15] is already implemented, there will be very little computation required as the edge detection has already been performed in it.

An edge map is created for each frame using Sobel operator. For a pixel at position (i, j) with value, $v_{i,j}, i \in 1,2,...,height, j \in 1,2,...,width$, in an image frame, the edge vector, $\vec{E}i,j = \{Ex_{i,j}, Ey_{i,j}\}$, is computed as follows.

$$Ex_{i,j} = v_{i-1,j+1} + 2 \times v_{i,j+1} + v_{i-1,j+1} - v_{i-1,j-1} - 2 \times v_{i,j-1} - v_{i+1,j-1} \quad (1)$$

$$Ey_{i,j} = v_{i+1,j-1} + 2 \times v_{i+1,j} + v_{i+1,j+1} - v_{i-1,j-1} - 2 \times v_{i-1,j} - v_{i-1,j+1} \quad (2)$$

Where $Ex_{i,j}$ and $Ey_{i,j}$ represent the degree of differences in vertical horizontal directions respectively. Next, the amplitude of the edge vector is computed as follows:

$$Amp\left(\vec{E}_{i,j}\right) = \left|Ex_{i,j}\right| + \left|Ey_{i,j}\right| \quad (3)$$

Homogeneity of an MB with size M×N, where M and N are the height and width of the MB respectively, depending on MB modes, is determined by using the amplitude of the edge vector in the MB using Equation (3). If the sum of the magnitudes of the edge vectors at all pixel locations in the MB is less than Thd_H, it is classified as homogeneous MB. Otherwise, it is non-homogeneous. The MB homogeneity threshold Thd_H is a preset parameter, and the MB homogeneity measure $H_{r,c}$ is set to value as follows:

$$H_{r,c} = \begin{cases} 1 & \sum\limits_{i,j \in M\,x\,N \text{ block}} Amp\left(\vec{E}_{i,j}\right) < Thd_H \\ 0 & \sum\limits_{i,j \in M\,x\,N \text{ block}} Amp\left(\vec{E}_{i,j}\right) < Thd_H \end{cases} \tag{4}$$

where r and c represent the row and column indices of the examined MB. According to the above equation (4), $H_{r,c} = 1$ indicates the M×N MB is homogeneous, and is non-homogeneous if $H_{r,c} = 0$.

3 Standstill and Homogeneity Based Fast Motion Estimation Algorithm

The whole idea of the proposed algorithm is to detect the static background present in the head-shoulder video sequences, and skip the motion search procedure once the MB is identified as static. Although there is already a fast motion estimation algorithm [1] included in the H.264 reference software, our algorithm does not conflict with it. In fact, it is designed to work collaboratively with the existing fast motion estimation algorithms, as most of these algorithms focus on the search strategies with different steps and search patterns in order to reduce the computation complexity. Whereas in our algorithm, in addition to the simplified search patterns for each every MB, those classified as static MBs are exempted from motion search. Because of the large portion of static background in the head-shoulder video sequences, it is justified to expect a large timesaving in motion estimation procedure, and therefore, reduction in total encoding time.

Figure 1 shows the flowchart of the proposed algorithm. The algorithm works as follows: in the INTER coding section of the whole encoding process, for each frame, its frame difference and edge profile will be calculated preceding the RDO calculation. During RDO computation, for each of the 7 modes (16×16, 16×8, 8×16, 8×8, 8×4, 4×8 and 4×4), the three conditions (illustrated in Figure 1) will be tested for satisfaction before the actual motion search is performed sequentially. Once any of the three conditions is satisfied, the motion search will be skipped, and the program proceeds to the next mode.

A typical binary image illustrating decision map after the three conditions test is shown in Figure 2. The shown frame (top) is the 3rd frame of News sequence (CIF), and the decision map (bottom) is for 8×8 mode of RDO computation. The white MBs represent those MBs that motion search is necessary, while the black ones are where motion search is skipped. Figure 3 is another example illustrating the situation in the 47th frame of sequence Grandma (QCIF), where the left is the original frame and the right is the decision map.

4 Experimental Results

The proposed fast motion estimation algorithm was implemented using JM8.4 encoder provided by JVT. Experiments were done when the fast motion estimation algorithm from JVT-F017 [1] is turned on and off. In both cases, we compare our proposed algorithm with the original motion search strategy. According to the specifications [16], the test conditions are as follows: 1) Baseline profile is used; 2) MV search range is ±32; 3) Hadamard transform is used; 4) optimization is enabled; 5) reference frame number is 5; 6) UVLC is enabled; 7) MV resolution is ¼ pixels; 8) GOP structure is IPPP; 9) number of frames is 150.

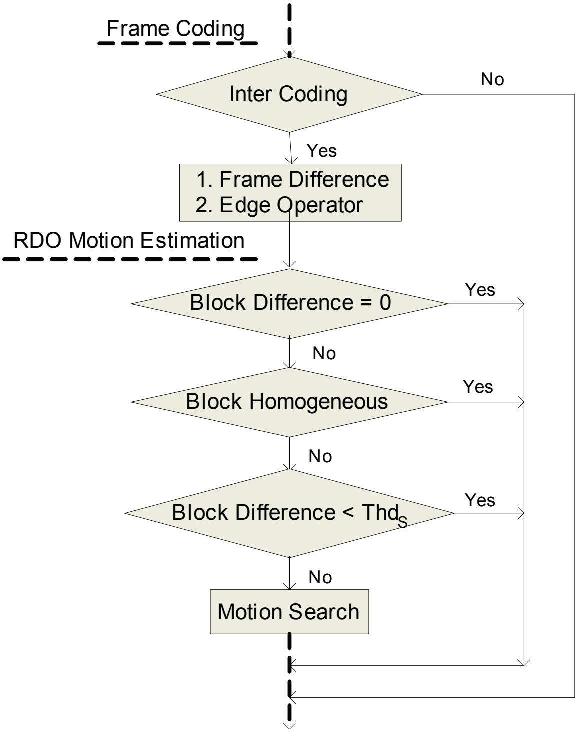


Fig. 1. Flowchart of the proposed algorithm

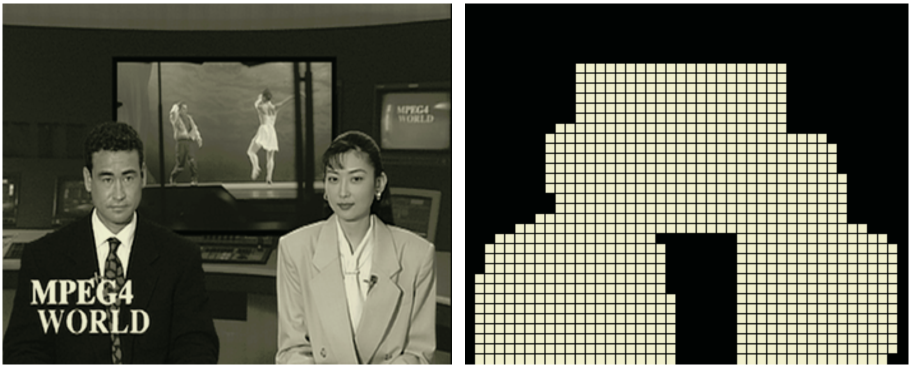


Fig. 2. Decision map for 3rd frame of News sequence

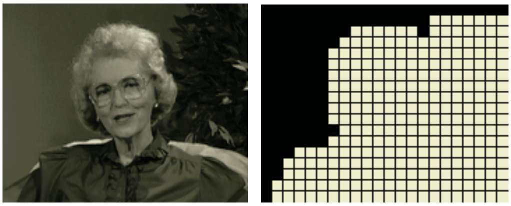


Fig. 3. Decision map for 47th frame of Grandma sequence

As our algorithm focuses on the head-shoulder sequences which are prevalent in the video-conferencing scenario, sequences like Akiyo, News, Silent, Grandma, and Mother and Daughter were chosen to carry out the experiments. In the experiments, the homogeneity threshold Thd_H is set to 20,000 for MB size of 16×16, and 20,000/(16×16/(M×N)) for MBs of size M×N. Likewise, Thd_S is set to be 200 for MB size of 16×16, and 200/(16×16/(M×N)) for MBs of size M×N. The results are tabulated in Table 1 and Table 2 corresponding to the cases where the fast motion estimation algorithm in [1] was switched off and on, respectively. It is noted that in the positive values in the tables represent increments, while negative values decrements.

From the experiment results, it is observed that the proposed approach has reduced the encoding time by 27% on the average. Consistent gain in coding speed for all the selected sequences in both cases, with least gain in the news video sequence and most gain in the grandma sequence. This is due to the larger stationary and homogeneous background present in the grandma sequence compared with that in news.

Table 1. Results with fast motion estimation [1] is off

Sequences	Format	Timesaving (%)	PSNR (dB)	Bitrate (%)
Akiyo	CIF	-25.65	-0.010	0.18
Mother	CIF	-15.23	-0.122	0.52
News	CIF	-11.38	-0.010	0.27
Paris	CIF	-12.03	-0.092	0.34
Grandma	QCIF	-30.82	-0.163	0.70
Silent	QCIF	-29.70	-0.081	0.44

Table 2. Results with fast motion estimation [1] is on

Sequences	Format	Timesaving (%)	PSNR (dB)	Bitrate (%)
Akiyo	CIF	-40.75	-0.020	0.19
Mother	CIF	-23.81	-0.110	0.47
News	CIF	-23.11	-0.015	0.29
Paris	CIF	-25.43	-0.174	0.49
Grandma	QCIF	-43.17	-0.200	0.71
Silent	QCIF	-37.35	-0.135	0.53

The overall consistency in the two cases in which the original fast motion estimation was switched on and off is also expected, since the proposed algorithm works among the existing motion estimation methods based on the standstill and homogeneity information.

5 Conclusions

A fast motion estimation algorithm which is able to work among the existing fast motion algorithms is proposed. The new method makes use of the temporal standstill characteristics as well as spatial homogeneity of the video sequences to decide whether a MB should go through the motion search procedure under each RDO mode decision. Frame difference together with a threshold value is used to judge whether the MB is time-stationary, and homogeneity decision is made based on the edge information. Therefore the time consuming motion estimation is skipped for standstill region, and is carried once for stationary region. The proposed algorithm is specially attractive to those head and shoulder sequences where real motion is confined to a small part of the frame. Experimental results show that it is able to achieve an average reduction of 27% encoding time, with a negligible average PSNR loss of 0.09 and 0.42% bit rate increase, for all the head and shoulder test sequences that we have tested.

References

1. Zhibo Chen, Yun He, et al. “Fast Integer and Fractional Pel Motion estimation,” JVT-F017.doc, 6th Meeting: Awaji, Island, JP, 5-13 December, 2002.
2. “Information technology-Coding of Audio-Visual Object – Part 10: Advanced Video Coding,” Final Draft International Standard, ISO/IEC FDIS 14496-10.
3. “Report of The Formal Verification Tests on AVC (ISO/IEC 14496-10 | ITU – T Rec. H.264),” MPEG2003/N6231, Dec. 2003, Waikoloa.
4. K.P. Lim, S. Wu, D.J. Wu, S. Rahardja, X. Lin, F. Pan, and Z. G. Li, “Fast Inter Mode Decision,” JVT-I020, 9th JVT meeting: San Diego, United States, September 2003.
5. J. R. Jain and A. K. Jain, “Displacement Measurement and Its Application in Interframe Image Coding”, IEEE Transactions on Communications, vol. 29, no. 12, pp.1799-1808, Dec. 1981.
6. T.Koga, K.Iinuma, A.Hirano, Y.Lijima, and T.Ishiguro, “Motion Compensated Interframe Coding for Video Conferencing,” in Proceedings Nat. Telecommunications Conf. 81, New Orleans, LA, pp.G5.3.1-G5.3.5, Nov. 1981.
7. L.M. Po and W.C. Ma, “A Novel Four-Step Search Algorithm for Fast Block Motion Estimation,” IEEE Transactions on Circuits and Systems for Video Technology, vol 6, No. 3, June 1996.
8. M. Chanbari, “The Cross-search Algorithm for Motion Estimation,” IEEE Transactions on Communications, vol. 38, No. 7, pp. 950-953, Jul 1990.
9. S. Zhu and K. K. Ma, “A New Diamond Search Algorithm for Fast Block Matching,” IEEE Transactions on Circuits and Systems on Video Technology, Vol. 9, No.2, pp. 287-290, Feb. 2000.
10. B. Liu and A. Zaccarin, “New Fast Algorithms for the Estimation of Block Motion Vectors”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 3, no. 2, pp.148-157, Apr. 1993.
11. Y. Kim, C.S. Rim, and B. Min, “A Block Matching Algorithm with 16:1 Subsampling and Its Hardware Design”, ISCAS’95, pp.613-616, 1995.
12. J. Feng, KT. Lo, H. Mehrpour, and A.E. Karbowiak, “Adaptive Block Matching Motion Estimation Algorithm Using Bit-Plane Matching”, IEEE International Conference on Image Processing, pp.496-499, 1995.

13. T. Uchiyama, N. Mukawa, and H. Kaneko, "Estimation of Homogeneous Regions for Segmentation of Textured Images," IEEE Proceedings in Pattern Recognition, pp. 1072-1075, 2000.
14. X. W. Liu, D. L. Liang, and A. Srivastava, "Image Segmentation Using Local Spectral Histograms", IEEE International Conference on Image Processing, pp. 70-73, 2001.
15. F. Pan, X. Lin, et al., "Fast Mode Decision Algorithm for Intra Prediction in JVT," JVT-G013, 7th JVT meeting, Pattaya, March 2003.
16. Gary Sullivan, "Recommended Simulation Common Conditions for H.26L Coding Efficiency Experiments on Low Resolution Progressive Scan Source Material," VCEG-N81, 14th meeting: Santa Barbara, USA. Sept. 2001.

Fast Rate-Distortion Optimization in H.264/AVC Video Coding

Feng Pan[1], Kenny Choo[2], and Thinh M. Le[2]

[1] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
efpan@i2r.a-star.edu.sg
[2] National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260
{eng10256,elelmt}@nus.edu.sg

Abstract. One of the new features in the H.264/AVC encoder is the use of Lagrangian Rate-Distortion Optimised (RDO) mode decision at the macroblock (MB) level. This brute-force algorithm searches through the 10 different MB coding modes to obtain the best one for encoding that MB, and is hence computationally expensive. This paper proposes a novel algorithm where the RDO can be reorganized in a better order such that the most likely MB modes will be tried first, and an early termination of the RDO process will be used once the calculated rate-distortion cost (RDcost) is below a preset threshold. The preset threshold is dependent on the RDcost of previous and neighbouring MBs that have been coded. This is based on the observation that the RDcost of an MB in the current frame is highly correlated to a co-located MB in the previous frame. Experimental results have shown that the new algorithm has dramatically reduced the encoding times of up to 61.8%, with negligible increases in bit-rates or PSNR.

1 Introduction

H.264/AVC is the newest international video coding standard [1]. It has been recently approved by ITU-T as Recommendation H.264 and by ISO/IEC as International Standard 14496–10 (MPEG-4 part 10) Advanced Video Coding (AVC). The elements common to all video coding standards are present in the H.264/AVC recommendation: MBs are 16×16 in size; luminance is represented with higher resolution than chrominance with 4:2:0 or 4:2:2 sub-sampling; motion compensation and block transforms are followed by scalar quantization and entropy coding; motion vectors are predicted from the median of the motion vectors of neighboring blocks; bidirectional B-pictures are supported that may be motion compensated from both temporally previous and subsequent pictures; and a direct mode exists for B-pictures in which both forward and backward motion vectors are derived from the motion vector of a co-located MB in a reference picture. Some new techniques, such as spatial prediction in Intra coding, adaptive block size motion compensation, 4×4 integer transformation, multiple reference pictures (up to 7 reference pictures) and content adaptive binary arithmetic coding (CABAC), are used in this standard. The test results of H.264/AVC show that it greatly outperforms existing video coding standards in both PSNR and visual quality [2].

An MB in H.264/AVC has 7 different motion-compensation block sizes in the Inter mode, 2 Intra prediction modes, and a SKIP mode so as to better adapt the size

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 433–441, 2005.

and shape of the blocks to the texture and motion of the video. In order to decide the best MB mode that could achieve the highest coding efficiency, H.264/AVC uses a non-normative technique called Lagrangian rate-distortion optimization (RDO) technique [3]. Figure 1 shows the RDO process employed in H.264/AVC. It is observed that RDcost values of all the possible modes for a given MB have to be computed so that a mode with the least RDcost value will be chosen as the best mode for coding this MD. Therefore, the computational burden of this type of brute force-searching algorithm is far more demanding than any existing video coding algorithm.

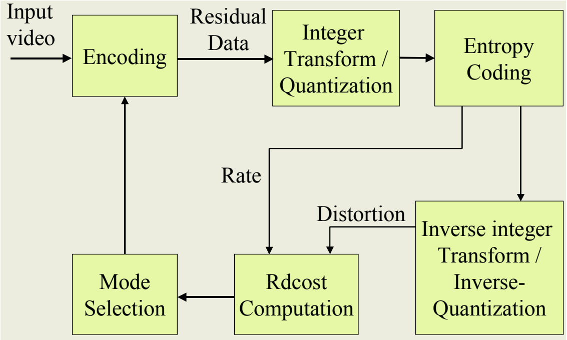


Fig. 1. Calculation of RDcost using brute force-searching

In general, a larger partition size usually suits slow motion and simple texture video objects, and a smaller partition size suits fast motion or complex scenes. Moreover, the probability of having different partition sizes in motion compensation is not uniform, and can be decided by using the information of the motion contents and spatial complexity, and in P-slices, the chance of having Intra modes is much less than that of Inter modes. Motivated by these observations, a number of efforts have been made to explore the fast algorithms in motion estimation, Intra mode prediction and Inter mode prediction for H.264/AVC video coding [4][6][7]. Fast motion estimation is a well-studied topic and is widely applied in the existing standards such as MPEG-1/2/4 and H.261/H.263. However, these fast motion estimation algorithms cannot be applied directly to H.264/AVC coding due to the variable block size motion estimation. On the other hand, fast mode decision is a new topic in H.264/AVC coding, and increasingly becomes an interesting research topic. It is believed that fast mode decision algorithms are also very important in reducing the overall complexity of H.264/AVC. A number of attempts have also been made to achieve fast mode decision for both Inter and Intra coding of MBs. In this paper, we presented a novel algorithm to achieve fast RDO in H.264 video coding. The proposed technique is based on the observation that the rate-distortion cost of an MB in the current frame is highly correlated to a co-located MB in the previous frame, and thus can be used as a means of early termination through thresholding. The presented algorithm considerably reduces the amount of calculations needed for Intra prediction with negligible loss of coding quality. The rest of the paper is organized as follows: Section 2 provides an overview of the current RDO algorithm, and the detailed description of the proposed algorithm; Section 3, summarized results of extensive simulations and Section 4, the conclusions.

2 Overview of Rate Distortion Optimization in H.264/AVC

As mentioned previously, H.264/AVC uses RDO technique to decide the best MB mode [3]. The RDO of H.264/AVC employs tree-structured inter coding by partitioning the MB into various sizes for motion estimation and compensation [4], and by using the directional prediction modes for Intra coding.

2.1 Encoding an MB by Different Block Partition and Directional Prediction

Generally speaking, in areas of spatial homogeneity and temporal stationarity, a large partition size (e.g. 16×16) would suffice; and in areas of high details, a smaller partition size would be appropriate (e.g. 8×8, 8×4, 4×8 or 4×4). Figure 2 shows all the possible MB modes for Inter coding. In addition to the above-mentioned 7 Inter Prediction modes, there are 3 other modes of SKIP, the Intra modes of I4MB and I16MB. SKIP mode is merely a direct copy of a co-sited MB in the previous frame. Intra Prediction [5] makes use of spatial correlation inside an MB to further remove the spatial redundancy at the MB level. H.264 defines 2 modes of Intra-prediction for luminance component, I4MB (4×4) and I16MB (16×16). In I4MB mode, there are 9 different prediction modes, of which 8 are directional prediction modes, and the 9th being DC mode. In I16MB mode, H.264 defines 4 predictions modes, of which 3 are directional and the 4th being DC mode.

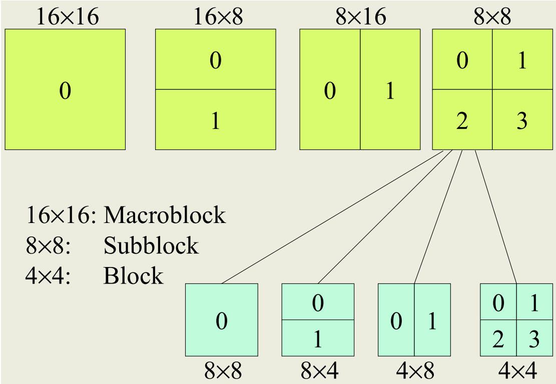


Fig. 2. Variable block size for rate distortion optimization

In order to choose the best block size for an MB, the H.264 standard makes use of computationally intensive RDO technique [2]. The general formula of RDO is given as:

$$J_{\text{mode}} = D + \lambda_{\text{mode}} \times R \tag{1}$$

where J_{mode} is the rate-distortion cost (RDcost) and $\square_{mode}$ is the Lagrangian multiplier; D is a measure of distortion between original and reconstructed MB, and R reflects the number of bits associated with choosing the mode and MB quantizer value, Q_p, including the bits for the MB header, the motion vector(s) and all the integer transform residue blocks. In Inter coding, possible modes are:

$$\text{mode} \in \left\{ \begin{matrix} \text{SKIP}, 16\times16, 16\times8, 8\times16, 8\times8, \\ 8\times4, 4\times8, 4\times4, \text{I4MB}, \text{I16MB} \end{matrix} \right\} \tag{2}$$

The optimal mode (mode decision) for an MB is selected that produces the least RDcost. The detailed steps of this brute-force full search RDO are as follows.

- Step 1. Perform motion estimation for all Inter modes
- Step 2. Compute MB RDcost, and compare the RDcost with the RDcost of the current best mode, and set the mode with the lower RDcost as the best mode.
- Step 3. If there are more valid modes, go to the next mode and go to Step 2, else go to Step 4
- Step 4. The mode with the lowest RDcost is selected as the mode to be encoded in. Go to next MB to encode and go to Step 1

It can be seen from the above steps that this algorithm performs motion estimation for all modes before comparing the RDcost. This is not necessary if the mode can be decided earlier via an early termination strategy through the use of contextual information of the video. Also, if the video objects contain much detailed motion, it is highly possible to be coded at small partition such as P8×8 mode. However the RDcost computation will still be carried on for all other modes, and only to come up with the same result – that the P8×8 mode is the best mode for encoding. This procedure is considerably time-consuming in performing needless computation in a bid for optimisation.

3 Fast Mode Decision of RDO

If the best mode can be determined at early stage of RDcost computation, significant timesaving can be achieved. The early termination strategy can be fulfilled based on the analysis of the video context, and by using empirical thresholds to shorten the computational process once a target threshold is achieved. In addition, motion estimation for any mode is performed only if there is a need to calculate the RDcost of this mode, and thus, the overall structure of the RDO process has to be modified to facilitate early termination.

3.1 Statistics of Mode Partitions in Test Video Sequences

In encoding of a natural video sequences, MBs in a slow-motion and low complexity sequence are usually coded using larger partitions such as SKIP or 16×16; whereas MBs in a fast-motion or high-complexity sequence are coded using smaller partitions such as 8×8, 8×4, 4×8 or 4×4 (P8×8). Because of the high correlation between consecutive frames in P-slices, the probability of encoding an MB using Inter mode is much higher than using Intra mode.

In order to verify the above observation, extensive experiments have been carried out to find the statistics of mode partitions in test video sequences. Figure 3 shows an example of the MB partition statistics of a number of test sequences by using full search RDO. It can be seen from the figure that for slow-motion QCIF sequences such as ‘Akiyo’, ‘Container’ and ‘Weather’, more than 85% of their MBs are encoded using the SKIP or 16×16 modes, whilst CIF sequences such as 'Paris' and 'Bike' have similarly high proportions of more than 65% of MBs encoded in SKIP or 16×16 mode. However, for fast-motion and high-complexity sequences such as 'Foreman' and 'Stefan', have almost 40% of their MBs are encoded using the other modes of smaller partitions. Therefore, significant timesaving can be achieved if we can design an intelligent early termination strategy during RDO by taking into account of the spatial and temporal complexity of sequence as well as the quantization values being used.

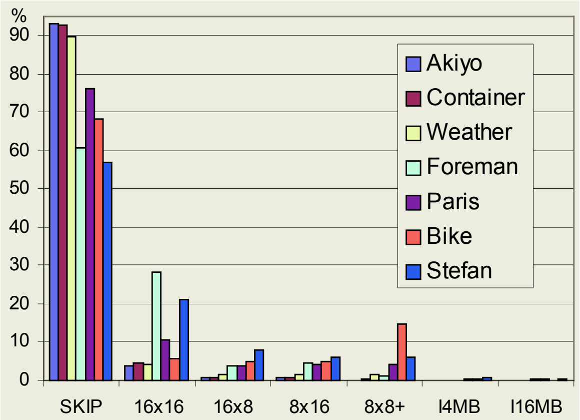


Fig. 3. Mode Distribution of 7 sequences for P-Frames, QP=40

3.2 P-Frame RDcost Correlation

The RDcost of MBs of subsequent frames shares a high degree of correlation. This is evident from experimentation, showing an average correlation of 0.988 for less complex sequences such as 'Akiyo', 0.939 for sequences such as 'Coastguard' and 0.972 for fast sequences such as 'Stefan'. This implies that RDcost provides a good basis for a threshold since the RDcost of the current MB is likely to be similar to the previous frame's co-located MB. Detailed correlation information is shown in Figure 4.

3.3 Proposed Algorithm

The first proposed idea in this paper is to restructure the RDO process into a more efficient structure. From Section 2.1 Step 1 and Figure 3, we have seen that the original algorithm performs motion estimation for all modes before proceeding to compute and compare the RDcosts. Computation due to motion estimation can be reduced significantly if motion estimation and RDcost computation for a mode are put together, such that thresholding will be able to save unnecessary computational time in performing motion estimation for the other thresholded modes.

In slow-motion sequences, temporal correlation is very high between frames, and therefore the RDcost and the mode of encoding of each matching MB in sequential frames should be similar to each other. Hence, the use of the RDcost of the previous frame's MB in the same position as a threshold is proposed. The threshold is given by Equation (3) below.

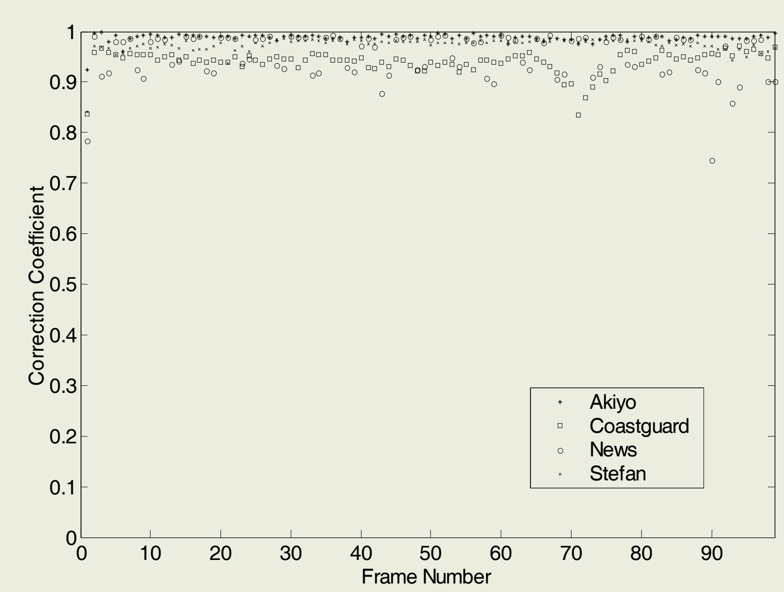


Fig. 4. Average RDcost correlation of four test Sequences

$$\theta_i^n = (1 \pm \alpha) \times C_i^{n-1} \tag{3}$$

where θ_i^n represents the threshold for i^{th} MB in current frame n, C_i^{n-1} is the actual RDcost of the co-located MB in the previous frame, α represents a restrictive factor, i is the index of the current MB in a frame, and n indicates the current frame, i.e.,

$$(1-\alpha) \times C_i^{n-1} \le \times C_i^n < (1+\alpha) \times C_i^{n-1} \tag{4}$$

The proposed algorithm is summarized in the following steps:

- Step 1. Perform motion estimation for the current mode if it is an Inter mode. Go to Step 2.
- Step 2. Compute the RDcost for the current mode. Go to Step 3.
- Step 3. If the RDcost for the current mode is less than the RDcost for the intermediate best mode, set the current mode as the intermediate best mode. Go to Step 4.
- Step 4. If the RDcost is within the threshold $\theta_{i,j}^n$, go to Step 6. Else go to Step 5.
- Step 5. If there are more available modes, go to the next mode and go to Step 1. Else go to Step 6.
- Step 6. Set the intermediate best mode as the best mode and use it for coding the MB.

The modified structure with the threshold uses the statistical analysis of sequences; hence the order of computation of modes starts with the SKIP and 16×16 modes, followed by a bit-rate minimizing strategy of 16×8, 8×16, P8×8 and the intra modes of I16MB and I4MB.

4 Experiments

4.1 Simulation Parameters

Our proposed algorithm was implemented into JM8.2 provided by JVT. According to the specifications provided in [12], the test conditions are as follows:

a. MV search range is ±32 pels.
b. Hadamard transform is used.
c. RD optimization is enabled.
d. Reference frame number equals to 5.
e. CABAC is enabled.
f. MV resolution is ¼ pel.
g. ABT mode is disabled.
h. GOP structure is IPPP.

A group of experiments were carried out on the recommended sequences with quantization parameters 28, 32, 36 and 40 as specified by [13] The comparison results were produced and tabulated based on the change of average data bits (ΔBitrate), the change of average coding time (ΔTime), and the change of average PSNR (ΔPSNR) [14].

4.2 Results

This section presents the simulation results of encoding in the proposed Fast RDO. All simulations were performed using a Pentium-4 3.06 GHz w/HT processor with 512 MB DDR RAM. The simulation results show that significant timesavings can be achieved through the Fast RDO method, which is especially effective in slow sequences such as ‘Akiyo’, ‘Container’ and ‘News’, with negligible changes in both bit-rate and PSNR. For more complex sequences, fast RDO is still able to achieve timesavings of up to 17.5% for ‘Coastguard’ with negligible loss of PSNR. The algorithm achieves on average - over all sequences tested - time savings of 34.01%, PSNR change of -0.038 dB and bit-rate change of 0.87%.

Table 1. Timesaving of the proposed algorithm

Sequences	Format	ΔTime (%)	ΔPSNR (dB)	ΔBit (%)
Akiyo	QCIF	-57.02	0.033	-0.562
Coastguard	QCIF	-17.52	-0.039	1.174
News	QCIF	-58.52	-0.049	0.907
Stefan	QCIF	-14.71	-0.120	2.490
Akiyo	CIF	-58.94	0.007	-0.096
Container	CIF	-39.85	0.031	-0.734
Foreman	CIF	-13.28	-0.072	1.617
Stefan	CIF	-12.25	-0.098	2.157

5 Conclusions

In this paper, we proposed a fast rate distortion optimization algorithm. Computation reduction was achieved due to the early termination of the full rate-distortion optimi-

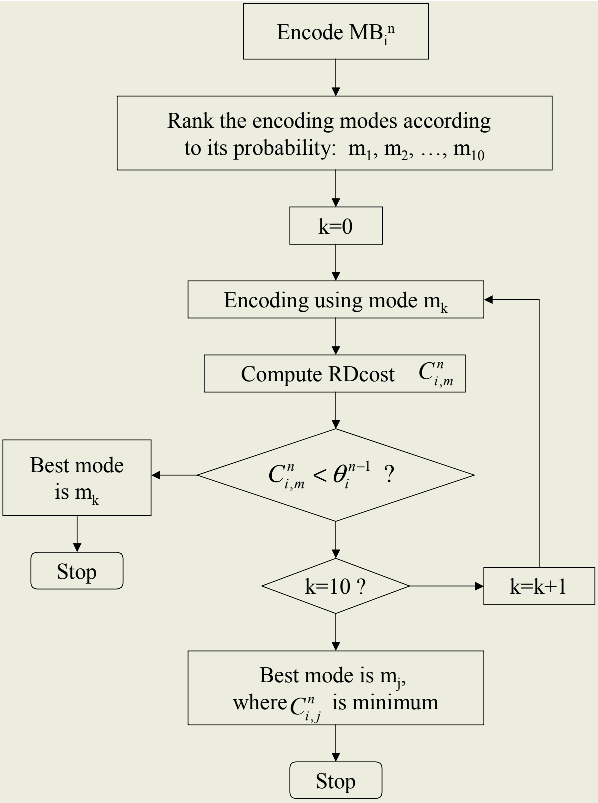

Encode MB_i^n
Rank the encoding modes according to its probability: $m_1, m_2, \ldots, m_{10}$
k=0
Encoding using mode m_k
Compute RDcost $C_{i,m}^n$
$C_{i,m}^n < \theta_i^{n-1}$?
Best mode is m_k
Stop
k=10 ?
k=k+1
Best mode is m_j, where $C_{i,j}^n$ is minimum
Stop

Fig. 5. Flowchart of the Proposed Fast RDO Algorithm

zation algorithm. This was through the use of a threshold which was proportional to the rate-distortion cost of the previously encoded frame. This exploited the high temporal correlation that slow-motion sequences have and achieved up to 58.94% timesavings in encoding slow sequences such as 'News' with negligible effects on the bitrate and PSNR. This algorithm is especially effective for slow and non-complex sequences. On faster or more complex sequences such as 'Stefan', the PSNR drop is higher, but still within acceptable norms. On the average, the proposed algorithm has reduced the encoding times by more than for slow sequences, and the timesaving for other sequences is also substantial.

References

1. "Information technology - Coding of audio-visual objects - Part 10: Advanced video coding," Final Draft International Standard, ISO/IEC FDIS 14496-10, Dec. 2003.
2. "Report of The Formal Verification Tests on AVC" (ISO/IEC 14496-10 | ITU-T Rec. H.264), MPEG/N6231, December 2003, Waikoloa, USA
3. Gary Sullivan, Thomas Wiegand and Keng-Pang Lim, "Joint Model Reference Encoding Methods and Decoding Concealment Methods,' JVT-I049d0, 9th JVT Meeting, San Diego, United States, September 2003.

4. Iain E. G. Richardson, “H.264/MPEG-4 Part 10 White Paper – Prediction of Macroblocks in P-Slices”, http://www.vcodex.fsnet.co.uk/h264_interpred.pdf
5. Iain E. G. Richardson, “H.264/MPEG-4 Part 10 White Paper – Prediction of Intra Macroblocks”, http://www.vcodex.fsnet.co.uk/h264_intrapred.pdf
6. Zhibo Chen, Peng Zhou, Yun He, “Fast Integer Pel and Fractional Pel Motion Estimation for JVT,” JVT-F017, 6th Meeting, Awaji Island, Japan, December 5-13, 2002.
7. Thinh M. Le, R. Mason, and S. Panchanathan, “Low Complexity Block Motion Estimation Using Morphological Image Processing and XOR Operations”, SPIE Journal of Electronic Imaging, vol. 09 (02), pp.110-116, Apr. 2000
8. F. Pan, X. Lin, S. Rahardja, K. P. Lim, Z. G. Li, G. N. Feng, D. J. Wu, and S. Wu, “Fast Mode Decision Algorithm for JVT Intra Prediction.” JVT-G013, 7th JVT Meeting, Pattaya II, Thailand, March 2003.
9. K. P. Lim, S. Wu, D. J. Wu, S. Rahardja, X. Lin, Feng Pan, and Z. G. Li, “Fast Inter Mode Decision.” JVT-I020, 9th JVT Meeting, San Diego, United States, September 2003.
10. Iain E. G. Richardson, “H.264 and MPEG-4 Video Compression – Video Coding for Next-Generation Multimedia”, John Wiley & Sons, 2003.
11. Mohamed Ghanbari, “Standard Codecs: Image Compression to Advanced Video Coding”, Institute of Electrical Engineers, London, UK, 2003.
12. Gary Sullivan, “Recommended Simulation Common Conditions for H.26L Coding Efficiency Experiments on Low Resolution Progressive Scan Source Material,” VCEG-N81, 14th meeting: Santa Barbara, CA, USA. September 24-27, 2001.
13. JVT Test Model Ad Hoc Group, “Evaluation Sheet for Motion Estimation,” Draft version 4, Feb. 19, 2003.
14. Gisle Bjontegaard, “Calculation of Average PSNR Differences between RD-curves,” VCEG-M33, 13th meeting: Austin, Texas, USA, April 2-4, 2001.

Improving Image Quality for JPEG Compression

Chin-Chen Chang[1,2], Yung-Chen Chou[2], and Jau-Ji Shen[3]

[1] Department of Information Engineering and Computer Science,
Feng Chia University, Taichung, Taiwan, 40724, R.O.C.
ccc@cs.ccu.edu.tw

[2] Department of Computer Science and Information Engineering,
National Chung Cheng University, Chiayi, Taiwan, 621, R.O.C.
jackjow@cs.ccu.edu.tw

[3] Department of Information Management,
National Formosa University, Yunlin, Taiwan, 632, R.O.C.
amitofo@sunws.nhust.edu.tw

Abstract. JPEG is a well-known method for digital image compression. The drawback of JPEG is distortion to image quality by truncation errors from quantizing the DCT coefficients. Another problem is the damage on the extracted watermarks from JPEG compression for the spatial-domain watermarks. In this paper, we proposed an improved method to enhance the quality of the decompressed image. The purpose of our proposed method is to adjust the reconstructed DCT coefficients closed as their original values. The coefficient adjustment blocks are used to guide the decompression operation during the coefficient reconstruction. Experimental results show that the proposed method has indeed made the image quality improvement as comparing to the images in the traditional method.

1 Introduction

Due to the digital technique advancement, a large number of images are being produced and processed daily. There are many important digital image issues being discussed the copyright protection. The invisible watermark techniques are used to defend rightful ownership of images as well as retaining the quality of watermarked images. The digital watermark techniques (or data hiding) are briefly classified into two categories for spatial and frequency domains, respectively. In the frequency domain techniques, watermark signals are inserted into the DCT or DFT coefficients transformed from the image. Whereas, the spatial domain techniques embed the watermark signals directly into the image pixels. The frequency-domain watermark maybe damaged during image compression such as the usual JPEG compression.

In order to reduce the large range of real numbers, using a quantization table is an easy way to achieve this goal. Discussions on the design of the quantization table can be found in [2][6][7][10]. According to the authors, a suitable quantization table should be effective for reducing the image size as well to preserve its quality. In 1999, Shohdohji et al. [8] applied the Fibonacci series to construct the quantization table. Meanwhile, Pan [3] used the different quantization tables to improve the image quality; and Park et al. [4] presented a high-resolution image reconstruction algorithm to modify the image data.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 442–448, 2005.

In order to keep the watermark quality after JPEG compression, we proposed a novel method to improve the decompressed image quality by doing a DCT coefficients' finely tuning. Our method can more precisely handle the rounding or truncations error due to the DCT coefficients' quantization.

This paper is organized as follows. In Section 2, discussions will be made on the JPEG procedures. In Section 3, we will propose a scheme to resolve the image distortion problem; further, a detailed discussion is followed. In Section 4 the performance of our proposed method will be discussed. Finally, the conclusions are made in Section 5.

2 Background

2.1 JPEG

The Joint Photographic Experts Group (JPEG) is a gray scale image compression standard developed in 1991 [9]. The DCT is the baseline for the JPEG image compression technique. The image compression system can be divided into two phases. The first phase is to compress an image. Initially, the image will be divided into $P \times Q$ non-overlapping blocks. Each block contains $M \times N$ (for instance 8×8) pixels. Then the DCT is applied to transform the blocks from spatial domain into frequency domain. The quantization mechanism is then applied to reduce the large range of the coefficient values in order to keep only the important and significant coefficients. After the quantization operation, Huffman coding is applied to encode the significant coefficients. The reduction in size of the compressed image is favorable for transmission, and less storage is required.

The second phase is to reconstruct the image from the compressed data on the receiver side. In the decompression stage, the entropy decoder is to recover the quantized coefficients from the compressed data, then quantization table is applied to recover the coefficients. Finally, the inverse DCT (IDCT) is applied to rebuild the image. However, the decompressed image is different from the original image, but not easily differentiated with raw human visual system.

2.2 Discrete Cosine Transform (DCT) and Zigzag Scan Ordering

The DCT was introduced in 1974 [1] to be used for transforming a digital image from its spatial domain into the frequency domain. The DCT transform is performed on a block unit containing $M \times N$ pixels. The following equations are used for defining FDCT as in Formula (1) and IDCT as in Formula (2), respectively.

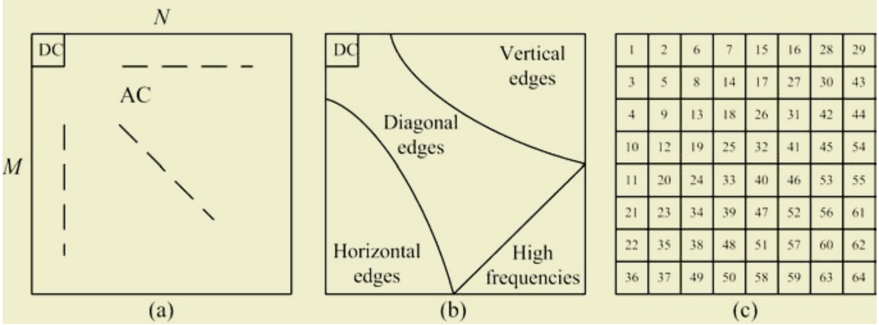
$$F(u,v) = \frac{1}{4}\alpha_u \alpha_v \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n) * \cos\left(\frac{(2m+1)\pi u}{2M}\right)\cos\left(\frac{(2n+1)\pi v}{2N}\right), \quad (1)$$

$$f(m,n) = \frac{1}{4}\sum_{u=0}^{U-1}\sum_{v=0}^{V-1} \alpha_u \alpha_v F(u,v) * \cos\left(\frac{(2m+1)\pi u}{2M}\right)\cos\left(\frac{(2n+1)\pi v}{2N}\right), \quad (2)$$

$$\alpha_u = \begin{cases} 1/\sqrt{2}, & u = 0 \\ 1, & 1 \le u \le M-1 \end{cases}, \text{ and} \quad (3)$$

$$\alpha_v = \begin{cases} 1/\sqrt{2}, & v = 0 \\ 1, & 1 \le v \le N-1 \end{cases} . \tag{4}$$

The notation *F*(*u*, *v*) is a coefficient value at coordinate (*u*, *v*). Here *f*(*m*, *n*) defines a pixel value at coordinate (*m*, *n*) in a block.



1	2	6	7	15	16	28	29
3	5	8	14	17	27	30	43
4	9	13	18	26	31	42	44
10	12	19	25	32	41	45	54
11	20	24	33	40	46	53	55
21	23	34	39	47	52	56	61
22	35	38	48	51	57	60	62
36	37	49	50	58	59	63	64

Fig. 1. (a) DC and AC coefficients distribution; (b) Structural decomposition of coefficients; (c) the zigzag scan order

A DCT block contains one direct current (DC) coefficient and *M*×*N*-1 alternating current (AC) coefficients (see Fig. 1(a)). In the AC area, the coefficients in the DCT block are decomposed into four different frequency levels [5] (see Fig. 1(b)). The coefficients closer to left top corner can be seen as more important coefficients thus the zigzag scan order (see Fig. 1(c)) is helpful for entropy coding.

3 The Proposed Method

Our method is designed to resolve the image distortion problem caused by JPEG compression as mentioned in Section 1. At the dequantization step, the DCT coefficients are adjusted in order to enhance the quality of decompressed image. Fig. 2 illustrates the flow for our proposed method. Following the image compression steps that have two phases. Two main parts are used to adjust the reconstructed coefficients for each phase; the purpose of which is to rebuild the coefficients to be as close as possible to the coefficient values in the original image. The first part is the coefficient adjustment block that is used to keep the information on the original coefficients with values modified during the quantizing operation. The second part is an adjustment bit map for the image, which is used to indicate the blocks which need the coefficient adjustment. After the adjustment information have been got, the entropy coder will encode the coefficient data, adjustment bit map and corresponding coefficient adjustment blocks.

3.1 Coefficient Adjustment Block

In order to reduce the impact of coefficient rounding or truncation error, the coefficient adjustment block is used to keep the information about whether a coefficient had undergone truncation, round off or neither of these two. Each element in the coefficient adjustment block can be presented by three individual states (0, 1 and 2).

Two bits are used to represent a state. State 0 indicates that a corresponding coefficient is either divided with no remainder or a too small remainder. In order to judge a small remainder, threshold TH_0 is needed. Therefore, the corresponding coefficients do not need to be adjusted. State 1 is used for quantized coefficients that have been rounded off. This is an indication to subtract a suitable adjustment value, which is percent (x) of corresponding quantization value, when the coefficient is recomputed. State 2 is used to indicate the quantized coefficient being truncated. State 2 implies to add a suitable adjustment value, which is percent (y) of the corresponding quantization value, when the coefficient is recomputed.

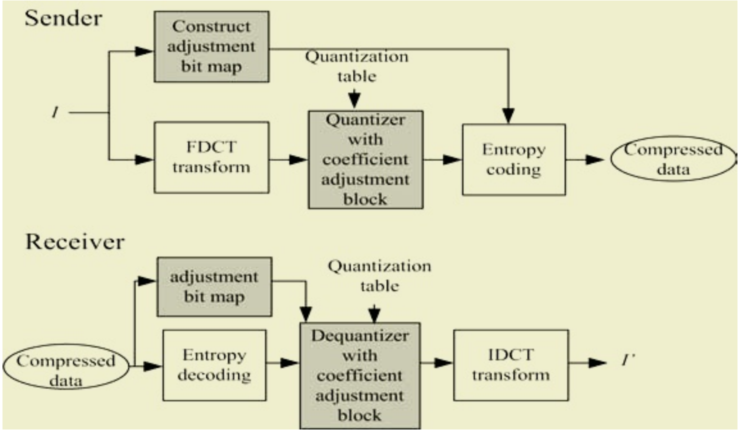


Fig. 2. Flow for the proposed method

After DCT, the important coefficients stayed in the left top part (see Fig. 1(a)). This means that the coefficient adjustment block does not need to keep the truncation and round off information for all of the coefficients in the quantized block. Therefore, a control parameter c is used to indicate the number of adjustment elements to be recorded in zigzag scan order.

3.2 Adjustment Bit Map

Observing the characteristic of images, a smooth block is unnecessary to adjust in its coefficients reconstruction. On the other hand, dropping the smooth block's adjustment information can reduce the extra bits requirement. In judging whether a block is a smooth block or not, we compute the variance of a block. Equation (5) is used to compute the variance of a spatial block.

$$\sigma = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\left|f(m,n)-\mu\right| \cdot \tag{5}$$

Here $f(m,n)$ is a pixel value in a block. The notation □ represents the mean pixel value in a block. A suitable threshold parameter TH_v is used to judge the type of a block. If the variance of a block is smaller than $TH_{v,}$ then the corresponding block is a smooth block. Equation (5) and TH_v are applied to setup a bit map to indicate smooth (0) and non-smooth (1) blocks of an image.

3.3 Reconstructing Coefficients

When the receiver receives a compressed image data, entropy decoder will decode this image. Its DCT coefficients will be reconstructed to recover this image. During the coefficient reconstruction, the adjustment bit map and coefficient adjustment blocks are applied to modify the recomputed coefficients such that their values are closer to their original ones. First, the adjustment bit map helps to decide whether the coefficient values of its corresponding coefficient block need to be adjusted or not. Next, if the adjustment of a block is necessary (i.e., its corresponding adjustment bit map entry is “0”.) then the adjustment information in the coefficient adjustment block provides the appropriate adjustment decisions to tune the coefficient values. If the adjustment state is 0, then there is no change on the corresponding coefficient value. If the adjustment state is 1 then the corresponding coefficient value should be subtracted x from its corresponding quantization value. If the adjustment state is 2 then the corresponding coefficient value should be added y on its corresponding quantization value.

4 Experimental Results

In order to test the performance of our proposed method, implementation is done by using the MATLAB software. The operating system is Windows XP with the hardware platform consisting of the Celeron 1.7GHz CPU and 512M RAM. The Peak Signal to Noise Ratios (PSNR) is used as the measure for image quality. The following equation defines the PSNR computation.

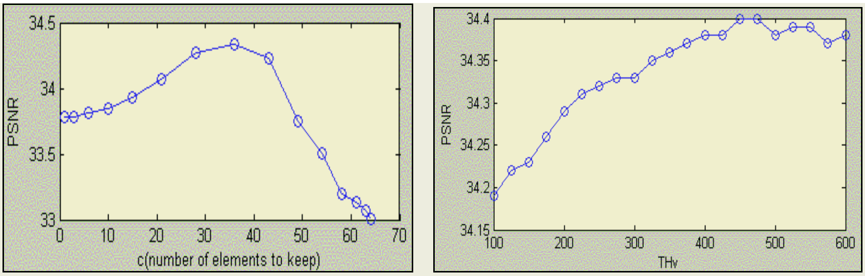
$$PSNR = 10\log_{10}\frac{255^2}{MSE} \tag{6}$$

$$MSE = \frac{1}{H \times W}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}(\beta_{ij} - \gamma_{ij})^2 , \tag{7}$$

where H and W stand for the image height and width, respectively. β_{ij} is the pixel value for image β at coordinates (i, j), and γ_{ij} is the pixel value for image γ with the coordinates (i,j). The Mean Square Error (MSE) calculates the difference between two images. Each block is 8×8 in its size. The parameter TH_v is set to 450. The parameters of x and y are set to 0.25 and 0.2, respectively. The control parameter c is set to 36. Comparisons are made on the performance between the traditional and the proposed methods. Fig. 3 (a) shows the results by using different c’s. The highest PSNR occurred when c is set to 36. Therefore, 36 is a suitable value for c. Fig. 3(b) shows different PSNRs from using different TH_v values. As can be seen, the TH_v value of 450 gives a higher *PSNR*. Table 1 illustrates results from the comparisons. It is obvious that the proposed method has higher PSNRs in the recovered images.

5 Conclusions

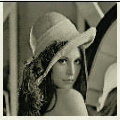
Although image compression serves the purpose of reducing the sizes of images, but the quality of these images should be preserved also. The frequency based image compression methods usually affect the quality of the decompressed images whose



(a) (b)

Fig. 3. (a) Experiments to determine a suitable *c*; (b) Experiments to determine a suitable *THv*

Table 1. Comparisons between the traditional method vs. the proposed method

	Traditional	Proposed method
Baboon (128×128)	 PSNR = 33.30	 PSNR = 34.27(need extra storage 2k)
Lena (128×128)	 PSNR = 33.82	 PSNR = 34.40(need extra storage 1.8k)
Pepper (128×128)	 PSNR = 34.07	 PSNR = 34.82 (need extra storage 1.8k)

results are caused by truncation of the coefficient values during quantization. Although our proposed method requires a small extra amount of storage to keep the information on truncation and round off manipulations, the efforts are worthwhile for improving the quality of the recovered image. Experimental results showed that the quality of the recovered images have been significantly improved. Thus, we can say that our proposed method is indeed feasible and efficient.

References

1. N. Ahmed, T. Natarajan and K. P. Rao, Discrete cosine transform, IEEE Transactions on Computers, Vol. C-23, (1974), 90-93
2. D. M. Monro and B. G. Sherlock, Optimum DCT quantization, Proceedings of Data Compression Conference, March & April (1993) 188-194
3. F. Pan, Adaptive image compression using local pattern information, Pattern Recognition Letters, Vol. 23, (2002) 1837-1845
4. S. C. Park, M. G. Kang, C. A. Segall, and A. K. Katsaggelos, Spatially adaptive high-resolution image reconstruction of dct-based compressed images, IEEE Transactions on Image Processing, Vol. 13, No. 4, (2004) 573-585

5. K. R. Rao and J. J. Hwang, Techniques and standards for image, video and audio coding, Upper Saddle River, N.J.: Prentice Hall, (1996)
6. V. Ratnakar and M. Livny, RD-OPT: An efficient algorithm for optimizing DCT quantization tables, Proceedings for Data Compression Conference, March (1995) 332-341
7. V. Ratnakar and M. Livny, An efficient algorithm for optimizing DCT quantization, IEEE Transactions on Image Processing, Vol. 9, No. 2, (2000) 267-270
8. T. Shohdohji, Y. Hoshino and N. Kutsuwada, Optimization of quantization table based on visual characteristics in DCT image coding, Computers and Mathematics with Applications, Vol. 37, (1999) 225-232
9. G. K. Wallace, The JPEG still picture compression standard, IEEE Transactions on Consumer Electronics, Vol. 38, No. 1, (1992) xviii-xxxiv
10. A. B. Watson, Visually optimal DCT quantization matrices for individual images, Proceedings of Data Compression Conference, March & April (1993) 178-187

Low-Power MPEG-4 Motion Estimator Design for Deep Sub-Micron Multimedia SoC*

Gyu-Sung Yeon, Chi-Hun Jun, Tae-Jin Hwang, Seongsoo Lee, and Jae-Kyung Wee

School of Electronics Engineering, Soongsil University, Seoul 156-743, Korea
wjk@ssu.ac.kr

Abstract. This paper proposes novel low-power MPEG-4 motion estimator with deep submicron technologies below 0.13μm. The proposed motion estimator reduces both dynamic and static power consumption so that it is suitable for large leakage process technologies. It exploits breaking-off search to reduce dynamic power consumption. To reduce static power consumption, block-wise shutdown method is employed. From the simulation results, power consumption was reduced to about 60%. The proposed motion estimator was designed in Verilog HDL and the estimated gate counts are about 45,000 gates.

1 Introduction

In SoC (System-on-Chip) design, one serious issues is power consumption, especially in the portable multimedia devices [1]. In the multimedia SoC, motion estimation [2] dominates total computation and power consumption. In MPEG-4 [3], motion estimation usually occupies more than 50% of total computation, so it is necessary to develop low-power motion estimator for multimedia SoC. Recently, several low-power motion estimators are proposed [4],[5]. However, they reduce only dynamic power consumption by exploiting low-computation algorithms, and they are not suitable for deep sub-micron process technology with large static power consumption.

This paper proposes novel low-power motion estimator architecture. It considers both dynamic and static power, while conventional architectures consider only dynamic power. It exploits three-step search (TSS) algorithm [6] for low computation. To reduce dynamic power, breaking-off search is adopted. Pipelined block-wise shutdown [7] to turn-off unused blocks is employed to reduce static power.

2 Power Reduction

(1) Dynamic Power Reduction: The proposed motion estimator exploits three-step search (TSS) [6] algorithm. It is widely used in mobile multimedia SoCs due to low computation and relatively simple hardware. Most of the motion estimation algorithms find the “most similar block” with the reference block. The sum of absolute difference (SAD) is used as a matching criterion of “most similar”, and the block with minimum SAD is regarded as the best-matching block. When the motion estimator matches the blocks in the search range, there is no possibility for the current matching

[*] This work was supported by the grant No. R01-2002-000-00039-0 from the Basic Research Program of the Korea Science & Engineering Foundation

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 449–455, 2005.

block to be the best-matching block if partially calculated value of SAD in the current matching block is larger than the minimum SAD of previous matches. In this case, the motion estimator needs no further calculation. It stops current matching block and skips to next matching block. This is called as "breaking-off search" (BOS) algorithm. When breaking-off occurs, it stops its operation during the remained calculation period for the current matching block and enters shutdown mode, thus dynamic power consumption is greatly reduced.

(2) Static Power Reduction: Block-wise shutdown [8] adaptively cut off the power supply of the function module when it idles. Large-size MOS cut-off switch is inserted between the power supply and the logic circuit. This scheme has simple hardware and easy control flow, and reduces both dynamic and static power consumption. However, it requires short wake-up time to shutdown blocks in time as operating frequency increases in deep sub-micron process, which causes large current surge during wake-up process. This often leads to system malfunction due to severe power line noise. To avoid this problem, pipelined block-wise shutdown [7] was proposed, where the block shuts down and wakes up sequentially along with pipeline stage. It successfully reduces current surge during wake-up process, since the number of simultaneous switching of logic gates is reduced. When the breaking-off occurs, the proposed motion estimator is shutdown, and the static power consumption becomes theoretically zero.

3 Implementation

Fig. 1 shows the architecture of the proposed motion estimator adopting the breaking-off search. The proposed motion estimator consists of a sliding window block, a difference block for obtaining SAD, a Wallace tree block, a comparator for performing a breaking-off processing, and a register for saving the minimum previous SAD value. For simplicity, half-pel search unit is omitted in Fig. 1. The proposed architecture can be easily applied to most fast block matching algorithms by modifying address generators. The sliding window block carries out the parallel calculation with 32bit registers. The address generator block generates the address of pixels used for next calculation, which is obtained by the position of the current macroblock and that of the reference macroblock. The proposed motion estimator exploits local memory to store search window data to reduce memory bandwidth.

Fig. 2 illustrates the proposed motion estimator with a pipelined wake-up circuitry. It consists of four parts: ME block, FIFO queue, shutdown controller, and state register. The ME block is pipelined with three partitions to reduce the power line noises coming from charging current during a block shutdown operation [7]. The block switching operates with cut-off PMOS switches between V_{DD} line and pipelined stages. FIFO queue saves the motion vector for macroblock grouping. (Macroblock grouping is described in Sect. 4 in detail.) Shutdown controller generates the signal for shutdown and wake-up. In the initial state, the value of FIFO's depth is set to grouping register and the number of stored motion vectors is equal to the depth of FIFO queue in group register. Group register sends the shutdown enable signal to shutdown controller. Shutdown controller sends the shutdown request signal to memory management unit block. Memory management unit writes the shutdown state in the state register and outputs a acknowledge signal to shutdown controller. It makes the delayed signals (wake-up/shutdown[j] in Fig. 2 where j=1,2,3) for pipeline stages with the received "acknowledge" signal as shown in Fig. 2. The delayed signals shutdowns the blocks with time delays in the pipeline stages. As shown in Fig. 2, the processing state in state register is stored during delay periods of the control signals. State register indicates shutdown state to memory management unit.

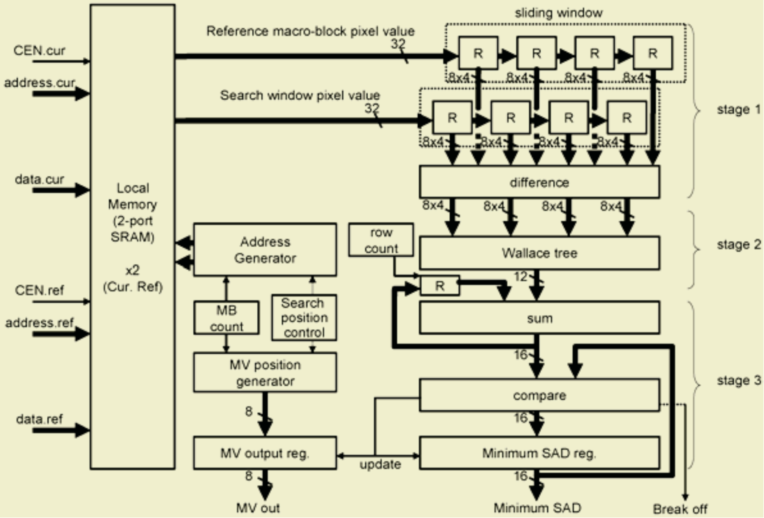


Fig. 1. Architecture of the proposed motion estimator

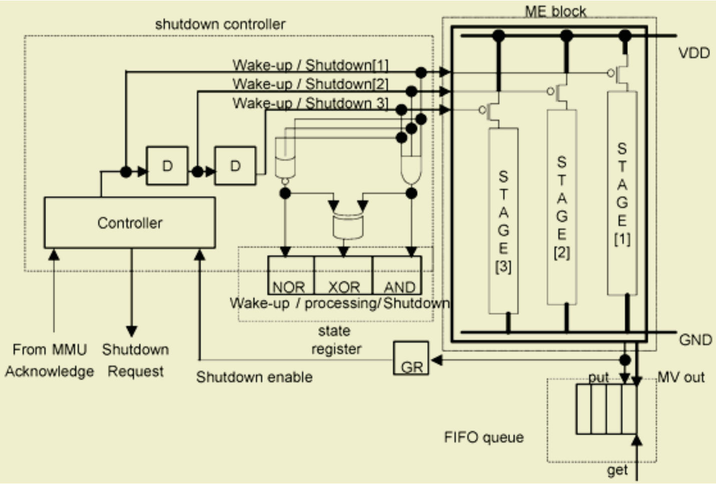
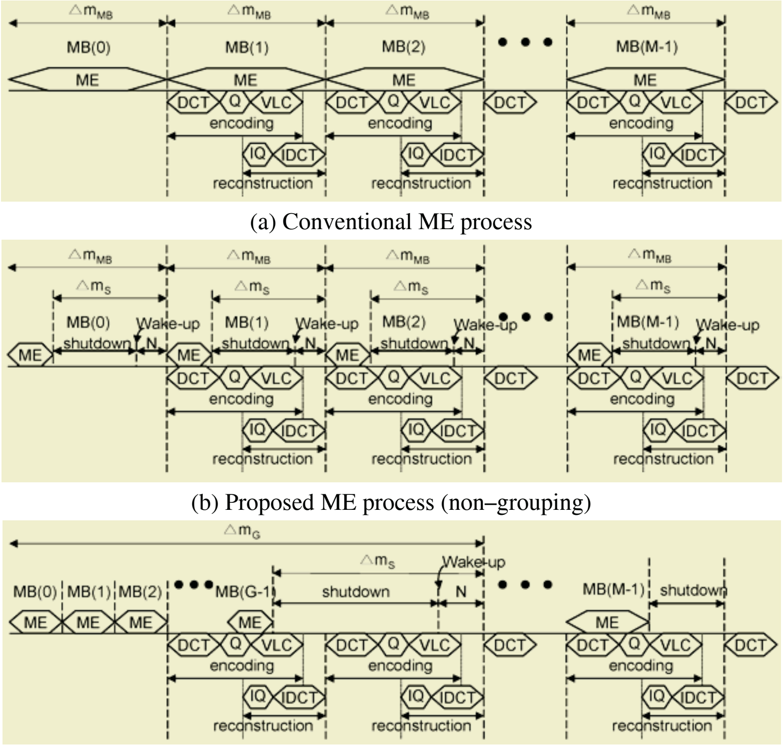


Fig. 2. Architecture of the pipelined wake-up in the proposed motion estimator

The proposed motion estimator fabricated with a 0.13μm MTCMOS technology was designed for mobile multimedia SoC with the search range of [-8, 8]. Core block of the motion estimator was about 40,000 gates including half-pel search unit. Shutdown controller, state register, and FIFO queue were about 5,000 gates.

4 System Level Control Method

In MPEG-4 codec, there are several major functional units such as motion estimation (ME), discrete cosine transform (DCT), inverse discrete cosine transform (IDCT), quantization (Q), inverse quantization (IQ), and variable length coding (VLC). For conventional MPEG-4 codecs, the execution diagram of these units is as shown in Fig. 3(a). The Δm_{MB} is the time assigned for each macroblock processing. There are three parallelized executions, i.e. motion estimation (ME), encoding (DCT-Q-VLC), and reconstruction (IQ-IDCT). Note that, in Fig. 3(a), motion estimator (ME) runs during Δm_{MB}, and it does not stop its operation at all. Fig. 3(b) shows the execution diagram of the proposed motion estimator. The Δm_S is the slack time by breaking-off. Operation time of the proposed motion estimator is (Δm_{MB} - Δm_S). The shutdown time is given by (Δm_S - 2N), where N is the number of pipeline stage. Note that Δm_S varies with each macroblock, since the actual workload of the motion estimation depends on the breaking-off status of each macroblock. To achieve further power reduction, we propose to group several macroblocks, as shown in Fig. 3(c). In the grouped macroblocks, the motion estimation is continuously performed without waiting Δm_{MB}. It increases slack time, since the number of shutdown and wake-up decreases. Consequently, it increases the efficiency of static power reduction.



(a) Conventional ME process

(b) Proposed ME process (non–grouping)

(c) Proposed ME process (grouping: G is grouping factor)

Fig. 3. Execution diagram of units

5 Simulation Results

We assumed that the dynamic power consumption is proportional to the execution cycle of the motion estimator, and no dynamic power is consumed when it idles. This is true in general for most video signal processing. When the break-off search is exploited, its execution cycle and corresponding dynamic power consumption are reduced. The normalized dynamic power consumption of the proposed motion estimator $P_{N,D}$ is given as Eqn. (1). It is simply the ratio of the execution cycles $N_{Break\text{-}Off}$ and $N_{NoBreak\text{-}Off}$, where $N_{Break\text{-}Off}$ and $N_{NoBreak\text{-}Off}$ are the execution cycles of the proposed and conventional motion estimators, respectively.

Static power consumption is proportional to the execution time of the motion estimator and no static power is consumed when it is shutdown. When the operating frequency is fixed, the execution time is proportional to the execution cycle. The normalized static power consumption of the proposed motion estimator $P_{N,S}$ is given as Eqn. (2). It is simply the ratio of the execution cycles ($N_{Break\text{-}Off}$ + $N_{Shutdown}$+ $N_{Wake\text{-}Up}$) and $N_{NoBreak\text{-}Off}$, where $N_{Shutdown}$ and $N_{Wake\text{-}Up}$ are the required execution cycles for shutdown and wake-up, respectively.

$$P_{N,D} = N_{Break-off} / N_{NoBreak-off} \tag{1}$$

$$P_{N,S} = (N_{Break-off} + N_{Shutdown} + N_{Wake-Up}) / N_{NoBreak-off} \tag{2}$$

$$P_{N,T} = P_{N,D} \times DPR + P_{N,S} \times SPR \tag{3}$$

In the SoC chip, the proportion of the dynamic power consumption and that of static power consumption to the total power consumption strongly depend on the fabrication process technology, and these values are quite constant over chip size. We call these values as dynamic power ratio (DPR) and static power ratio (SPR), respectively. Since the nano-scaled process technologies beyond 0.1 μm are not available to us, we modeled the normalized total power consumption of the proposed motion estimator $P_{N,T}$ as Eqn. (3). The values of DPR and SPR can be estimated in the several research articles, and they are shown in Table 1.

Table 2 shows the simulation results of the normalized dynamic power consumption $P_{N,D}$, the normalized static power consumption $P_{N,S}$, and the normalized total power consumption $P_{N,T}$ with 0.13 μm technology. The simulation was performed with CAD tools supported by IC Design Education Center. From the simulation results, the total power consumption is reduced by 61~70%. Note that the static power consumption is less reduced than the dynamic power consumption. This comes from the fact that the shutdown time of the proposed motion estimator is shorter than the idle time since extra cycles are needed during shutdown and wake-up.

Table 1. Ratio of dynamic power and static power with fabrication process

	Fabrication process generation (μm)			
	0.13	0.10	0.07	0.05
DPR	0.9	0.8	0.6	0.5
SPR	0.1	0.2	0.4	0.5

Table 2. Simulation results

Test sequences	$P_{N,D}$	$P_{N,S}$	$P_{N,T}$
Trevor	0.39	0.43	0.39
Carphone	0.32	0.36	0.34
Miss America	0.34	0.38	0.34
Flower Garden	0.31	0.34	0.30

Recently, the static power consumption becomes a serious problem in the multimedia SoCs. Heat dissipation of large-size multimedia SoC increases the temperature in chip core. This increases the leakage current and the corresponding static power consumption. They also increase as the fabrication process technology progresses. The proposed scheme reduces both the dynamic power and the static power, while conventional schemes reduce only the dynamic power. Consequently, the power reduction of conventional schemes degrades as the proportion of the static power increases, while that of the proposed scheme is almost insensitive.

Fig. 4 shows the normalized power reduction vs. chip core temperature. As illustrated in Fig. 4, power efficiency of conventional low-power technologies shows severe degradation, while the proposed scheme is insensitive. Fig. 5 shows the normalized power reduction vs. fabrication process technology generation. Similar tendency as Fig. 5 can be also found in Fig. 5.

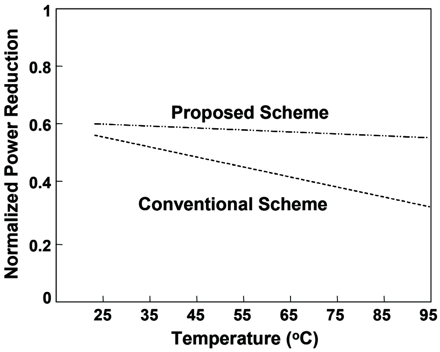


Fig. 4. Reduced total power consumption vs. temperature

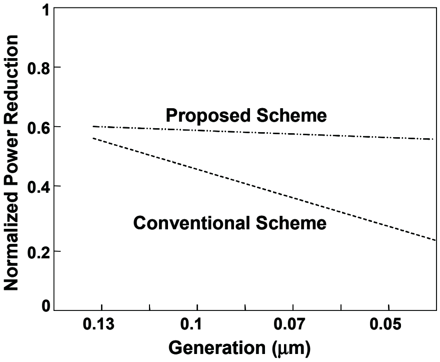


Fig. 5. Reduced total power consumption vs. process generation

6 Conclusion

This paper proposed a low-power motion estimator for multimedia SoC with deep sub-micron technologies below 0.13μm. It exploits breaking-off search to reduce the dynamic power consumption and block-wise shutdown to reduce the static power consumption. Its architecture is flexible and can be easily adapted for many motion estimation algorithms. It was designed in Verilog HDL, and the estimated gate count is about 45,000 gates. From the simulation results, the power consumption is reduced by about 60~70%. As the chip core temperature increases and the fabrication process technology progresses, the power reduction of the conventional scheme gets worse, while that of the proposed scheme is almost insensitive.

References

1. Chandrakasan, A., Brodersen, R.: Low Power Digital CMOS Design, Kluwer Academic Publishers (1995).
2. Jain, J., Jain, A.: Displacement Measurement and Its Application in Interframe Image Coding, IEEE Transactions on Communications **29** (1981) 1799-1808.
3. ISO/IEC JTC1/SC29/WG11 14496-2, Coding of Audiovisual Object: Visual (1998).
4. Richmond, R., Ha, D.: A Low-Power Motion Estimation Block for Low Bit-Rate Wireless Video, Proceedings of International Symposium on Low-Power Electronics and Design (2001) 60-63.
5. Kuroda, Y., Miyakoshi, J., Miyama, M., Imamura, K., Hashimoto, H., Yoshimoto, M.: A Sub-mW MPEG-4 Motion Estimator Processor Core for Mobile Video Application, Proceedings of the Asia South-Pacific Design Automation Conference (2004) 527-528.
6. Koga, T., Linuma, K., Hirano, K., Iijima, Y., Ishiguro, T.: Motion compensated interframe coding for video conferencing, Proceedings of National Telecommunication Conference (1981) G.5.3.1-5.3.5.
7. Choi, J., Kim, Y., Wee, J., Lee, S.: Pipelined Wake-Up Scheme to Reduce Power line Noise for Block-Wise Shutdown of Low-Power VLSI Systems, IEICE Transactions on Electronics **87** (2004) 629 - 633.
8. Shimizu, T., Arakawa, F., Kawahara, T.: Autonomous Decentralized Low-power System LSI Using Self-Instructing Predictive Shutdown Method, Proceedings of Symposium on VLSI Circuits (2001) 55 - 56.

Real-Time 3D Artistic Rendering System

Tong-Yee Lee, Shaur-Uei Yan, Yong-Nien Chen, and Ming-Te Chi

Department of Computer Science and Information Engineering,
National Cheng-Kung University, Tainan, Taiwan, Republic of China
tonylee@mail.ncku.edu.tw

Abstract. This paper presents an artistic rendering system for generating 3D images of Chinese paintings using graphics hardware. The user can adjust suitable parameters flexibly to generate different brush styles as his/her hobby, and see rendering results in real time. In this system, we propose a hardware-accelerated method to draw Chinese painting strokes efficiently along visible silhouettes. Three-dimensional texture and multi-texture from normal graphics hardware is used to speed up generating various brushes with Chinese painting stylized strokes. The features of the traditional Chinese painting such as ink diffusion and moisture effects are simulated. Several examples of aesthetically pleasing Chinese-paintings rendered from 3D models are demonstrated using the proposed method.

1 Introduction

In the past, most non-photo-realistic rendering (NPR) researches focus on the western painting styles such as pen-and-ink, watercolor, hatching and so on. However, few works in NPR are about Chinese paintings and most of them focus on simulating delicate effects of brush, black ink and papers. Furthermore, most of them are 2D Chinese drawing works and computationally expensive for real-time applications. These researches are interested in its simulated quality rather than in its processing time. However, when generating the scene of Chinese painting style in games or virtual environment, the real-time performance is required and we cannot use these previous works directly. In this paper, we present a real time NPR system for generating 3D Chinese paintings. The system pipeline consists of four stages and it is illustrated in Fig. 1.

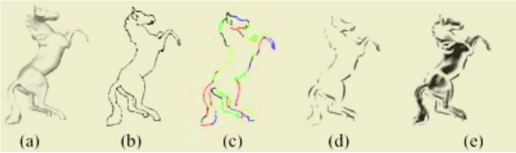


Fig. 1. System overview. (a) Input model, (b) visibility testing, (c) visible segment linking, (d) stroke placement and (e) interior shading. In (c), we color each linked segment

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 456–462, 2005.

2 Related Work

Strassmann models hairy brushes in his 2D oriental black-ink painting system [1]. This work represents each stroke with a cubic spline and renders the stroke using polygons with texture. Lee [2] designs a 3D brush model with elastic bristles that respond elastically to the force exerted by an artist against the paper. To simulate realistic diffuse effects of blank-ink paintings, Guo et al. consider the sorbency of paper, the liquid density and flow [3]. Zhang et al. [4] propose to use cellular automation-based simulation of ink behavior to render 3D trees in sumie style. Way et al [5] propose a method of synthesizing rock texture in Chinese landscape painting. Later, they further develop methods to simulate 3D Chinese painting trees using silhouettes with texture strokes [6]. Chan et al. [7] exploit existing software packages such as Maya and RenderMan to create 3D Chinese painting animation. Chu et al. [8] develop a system utilizing Haptic input device to track 3D brush movement to give more accurate brush simulation. Yeh et al. [9] propose a set of algorithms to render 3D animal models in Chinese painting style. To shade the interiors of animals, several basic image processing techniques such as color quantization, ink diffusion and box filtering are used.

3 Stylizing Silhouettes with Brush Strokes

3.1 Stroke Paths and Widths Generation

The idea for drawing view-dependent silhouettes of 3D model with stylized strokes is popular in NPR. We adopt Isenberg et al.' approach [10] to find visible silhouettes and concatenate silhouette segments into long stroke paths. Before applying stylizations onto stroke paths, control points along paths need to be interpolated using cubic spline to smooth the curvature of stroke paths. To make a stylized stroke path, we need to grow various widths at control points. In traditional Chinese painting, the brush width starts with thin stroke and gradually grows to thick stroke, and turns back to thin stroke as the brush stroke progresses. Yeh et al. [9] assign stroke width based on the order of control points only and potentially generate less smooth transition between different brush stroke widths when a long brush path contains fewer control points. To solve this problem, we consider distance between control points as another parameter to control the width of brush stroke. We use Eq. (1) and (2) to compute the width. In Eq. (1), ncp represents the total number of control points on a given stroke path. Eq. (2) represents the width of the stroke at a given control point i, which is 0 at both starting and ending points of the brush path. Eq. (1) returns the width to add or to subtract given the condition at the ith control point, where Vlength[i] is distance between the ith and the (i-1)th control points and width_step is a predetermined width step value. We demonstrate an example to compare [9] and our approach in Fig. 2. The proposed approach yields better visual stroke appearance than [9].

$$add[i] = \begin{cases} +VLength[i] * width\_step, & if \;\; i < \frac{1}{2} \times ncp \;\; and \;\; width[i] < MAX\_WIDTH \\ -VLength[i] * width\_step, & if \;\; i \geq \frac{1}{2} \times ncp \;\; and \;\; width[i] \geq 0 \\ 0, & else \end{cases} \tag{1}$$

$$width[i] = \begin{cases} 0 & ,if \;\; i = 0 \;\; or \;\; i = ncp \\ width[i-1] + add[i] & ,else \end{cases} \tag{2}$$

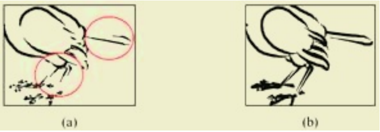


Fig. 2. Different stroke widths generated by different methods. Left: stroke generated by [9] Right: stroke generated by our system

3.2 Brush-Strokes in Chinese Painting Style

A brush consists of many bristles. In a microscopic view, when a single bristle draws on the paper, the effects it can produce different ink shades such are dark, light, dry and wet when ink diluted with water; with different pressure, direction in brush stroke, different ink tones can produce millions of variations of touches on the paper. Various brush-strokes are usually used to represent different texture of the subjects in Chinese painting. To simulate different shades of ink, we define the term ”pattern” to refer to the ink traces on the paper left by a bristle. The pattern setup is illustrated in Fig. 3 by combining an intensity map and an opacity (i.e., alpha) map. We can use 2D texture to store each pattern. When painting an absorbent paper, an artist can control the water content in the brush to make ink look sear, soggy or wet. The opacity map is used to control water content and therefore it is called a moisture map, too. The intensity map is used to control ink shades such as dark and light. With different combinations of bristle patterns, different style of brush strokes can be produced; Fig. 4 shows example of our simulated brush strokes in the style of flying-white (fei-bei) technique and slanted brush technique in Chinese painting. To efficiently generate the brush stroke patterns in real time, the hardware-accelerated 3D volume texture and multi-texture techniques are used in our system. Intensity changes are mapped to a 3D texture level to encode intensity change into the

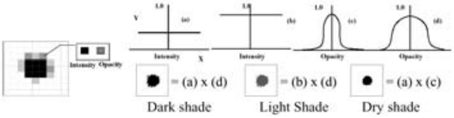


Fig. 3. Brush stroke setup on paper

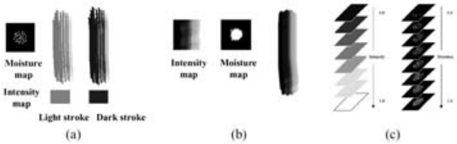


Fig. 4. (a) flying-white (fei-bei) brush path (Left). (b) slanted brush stroke path. (c) 3D texture representing intensity and moisture maps for brush strokes

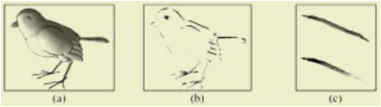
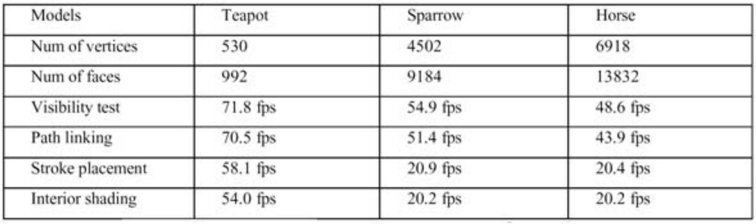


Fig. 5. By placing light, stroke brightness distribution can be controlled(Middle). Brush moisture effect(Left)

third dimension of 3D texture. Moisture changes are also mapped to another 3D texture level to account the moisture change in the third dimension as well. See figure 4(c) for a visual representation of this idea. The user just needs to prepare two sets of predetermined intensity and moisture maps. Then, we load these two sets into 3D volume texture. At running time, when an arbitrary intensity value is specified, corresponding 2D texture pattern can be interpolated efficiently from two neighboring intensity value automatically by the graphics hardware, thus a intensity map of the brush at the given intensity value is obtained. Similarly, the moisture map of a given moisture value can be computed in this hardware accelerated manner. With OpenGL extension [11], it allows us to enable multi-texture technique to combine the two maps into a new brush pattern. In this way, the system can deliver brush pattern with desired intensity and moisture value very fast using hardware accelerated 3D volume texture and multi-texture techniques. Next, we will give details about how to provide both intensity and moisture value at running time.

In Section 3.1, control points along every stroke path contain its original 3D coordinates and normal vector. Given a light source, we can compute a light vector from a control point to a light source. The intensity value of each control point is computed by the dot product of these two vectors and this dot product is normalized to the range of (0,1), so that all paths have different intensity values and are influenced by the lighting condition; distribution of bright and dark stroke can be gathered by giving different configuration to lighting as illustrated in Fig. 5. When an artist actually draws on paper, the moisture of the brush changes from wet to dry from the beginning of a stroke to the end of it. In order to simulate this effect, the control points in the beginning of each path are assigned a predetermined amount of moisture, and the moisture level in consequent control points drops as the distance from the first control point increases. See Fig. 5(c) for an example of the moisture effect. After the intensity and moisture value are found at each point along a brush path, we can use them as the third dimension of 3D volume texture to fast compute corresponding intensity and moisture stroke "pattern".

Table 1. Performance breakdown for Fig. 6



Models	Teapot	Sparrow	Horse
Num of vertices	530	4502	6918
Num of faces	992	9184	13832
Visibility test	71.8 fps	54.9 fps	48.6 fps
Path linking	70.5 fps	51.4 fps	43.9 fps
Stroke placement	58.1 fps	20.9 fps	20.4 fps
Interior shading	54.0 fps	20.2 fps	20.2 fps

4 Interior Shading

In this section, we present a method to draw colors in the interior area of models. The goal of this method is to fast simulate ink color change from dark to light like ink diffusion in Chinese painting. We use Eq. (3) to simulate this change in the interior of models.

$$Opacity = A * \cos^{n}(\frac{\theta}{W}) \tag{3}$$

Where θ is the angle between a vertex normal and a light vector from a vertex to light, A, n and W are constants to control shape of this function. To implement this simulation, the fragment shader provided by Nvidia's CG language [12] is used to do per pixel opacity value calculation using Eq. (3). By using the calculated opacity values and user-defined ink intensity for shading the object, the change of brightness can be seen after we enable blending. The brightness change can be seen as ink diffusion in Chinese ink painting. To avoid the dull appearance of uniform color distribution, noise functions such as Perlin noise can be used to add some randomness. In the next Section, several interesting results will be demonstrated to verify the proposed method for interior shading. In contrast to other ink diffusion approach [3, 4], the proposed method computes very fast but it yields not bad results.

5 Experimental Results

The experimental setup in this paper is a program written in C++ and OpenGL using Microsoft Visual C++ 6.0 compiler running on an Intel Pentium 4 R 2.2Ghz machine with Microsoft Windows 2000. Graphics card is Nvidia GeForce FX5900 with 256MB frame buffer. In Table 1, three models are used to test rendering speed measured in frame per second (fps) at 800x 600 screen resolutions. Because our system is implemented in four different stages, we list the frame rate (i.e., fps) at different stages. We can see the current performance bottleneck is limited by the stroke placement stage. The rendering performance we achieve is fast enough for user interaction in real time. Fig. 6 shows rendered results for Table 1. More results are demonstrated in Fig. 7.

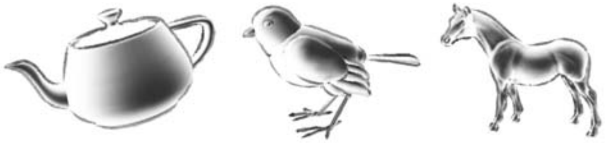


Fig. 6. We show three rendered results used in Table 1

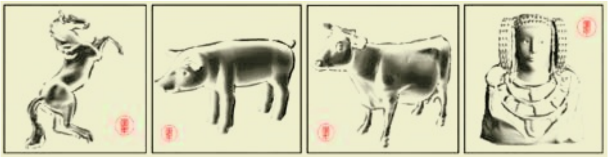


Fig. 7. More rendered results

6 Conclusion and Future Work

This paper presents a real-time NPR rendering system to generate traditional Chinese ink paintings for 3D models. The proposed method is accelerated by the normal graphics hardware. This method consists of four main steps: 1) visibility testing, 2) path linking, 3) stroke placement and 4) interior shading. For the stroke placement, the 3D volume texture and multi-texture techniques are used to fast compute ink and moisture information. We also attempt to simulate ink diffusion effect to paint the interiors of the models. As a result, many aesthetically pleasing Chinese-paintings rendered in real-time from 3D models are demonstrated using the proposed method. There are many possible future work can be further explored based on our current work. For example, we plan to consider motion issue in NPR such as sparrow jumping or waving its wings created by using bone and skin deformation techniques. In this situation, we need to consider the stroke coherence problem. Without treating well this issue, it is very easy to create popping effect during animation or deformation. Another our research direction is to how to express model metamorphosis [13, 14] or human face [15] in Chinese painting style or in western painting style [16].

Acknowledgement

This paper is supported by the National Science Council, Taiwan, Republic of China, under contract No. NSC-93-2213-E-006-026 and NSC-93-2213-E-006-060.

References

1. Strassmann, S., "Hairy brushes," Proc. SIGGRAPH 86, 20(4): 225-232, August 1986.
2. J. Lee, "Simulating oriental black-ink painting," Computer Graphics and Applications, IEEE, vol. 19(3), pp. 74-81, May-June 1999.

3. Q. Guo and T. Kunii, "Modeling the Diffuse Painting of Sumie'," Modeling in Computer Graphics (Proc. IFIP WG5.10), T. Kunii, ed., Springer-Verlag, Tokyo, 1991,pp. 329-338.
4. Q. Zhang, Y. Sato, J. Takahashi, K. Muraoka and N. Chiba, "Simple cellular automation-based simulation of ink behavior and its application to Suibokuga-like 3D rendering of trees," Journal of Visualization and Computer Animation, 1999.
5. Way, D. L., and Shih, Z. C. "The Synthesis of Rock Texture in Chinese Landscape Painting," Computer Graphics Forum, Vol. 20, No. 3, pp. C123-C131, 2001.
6. Way, D. L., Lin, Y. R. and Shih, Z. C. "The Synthesis of Trees Chinese Landscape Painting Using Silhouette and Texture Strokes." Journal of WSCG, Vol. 10, No. 2, pp.499-506, 2002.
7. C. Chan, E. Akleman, and J. Chen, "Two methods for creating Chinese painting," Proceedings.10th Pacific Conference on, October 2002, pp. 403 - 412.
8. N. S.-H. Chu and C.-L. Tai, "An efficient brush model for physically based 3d painting," in Proceedings of 10th Pacific Conference on Computer Graphics and Applications, 2002, 2002, pp. 413-421.
9. Jun-Wei Yeh, Ming Ouhyoung, "Non-Photorealistic Rendering in Chinese Painting of Animals," Journal of System Simulation, Vol. 14, No. 6, 2002, pp. 1220-1224.
10. T. Isenberg, N. Halper, and T. Strothotte, "Stylizing silhouettes at interactive rates: From silhouette edges to silhouette strokes," in Computer Graphics Forum, Proceedings of Eurographics 2002, vol. 21, September 2002, pp. 249-258.
11. http://www.opengl.org
12. http://developer.nvidia.com/page/cg_main.html
13. Tong-Yee Lee, P.H Huang, "Fast and Institutive Polyhedra Morphing Using SMCC Mesh Merging Scheme," IEEE Transactions on Visualization and Computer Graphics, Vol. 9, No. 1, pp. 85-98, 2003.
14. Chao-Hung Lin, Tong-Yee Lee, "Metamorphosis of 3D Polyhedral Models Using Progressive Connectivity Transformations," IEEE Transactions on Visualization and Computer Graphics, Jan./Feb. Issue, Vol. 11, No.1, pp. 2-12, 2005
15. Tong-Yee Lee, Ping-Hsien Lin, Tz-Hsien Yang, "Photo-realistic 3D Head Modeling Using Multi-view Images," in Lecture Notes on Computer Science (LNCS) 3044, Springer-Verlag, pp. 713-720, May 2004.
16. Ming-Te Chi, Tong-Yee Lee, "Stylized and Abstract Painterly Rendering System Using a Multi-Scale Segmented Sphere Hierarchy," to appear in IEEE Transactions on Visualization and Computer Graphics.

Index LOCO-I: A Hybrid Method of Data Hiding and Image Compression

Wei-Soul Li[1], Wen-Shyong Hsieh[1,2], and Ming-Hong Sun[1]

[1] Department of Computer Science and Information Engineering,
National Sun Yat-Sen University, NO. 70 Lien-hai Rd., Kaohsiung 804, Taiwan, R.O.C.
leews@mail.cse.nsysu.edu.tw
[2] Department of Computer Science and Information Engineering, Shu Te University,
Kaohsiung 824, Taiwan, R.O.C.
wshsieh@mail.stu.edu.tw

Abstract. This paper proposes an approach for the combined data hiding scheme with LOCO-I/JPEG-LS image compression algorithm. We present a new embedded image coder and investigate the lossless and near-lossless performance of these transforms in the propose coder. We will present our proposed encoder/decoder scheme of SNR scalability layer. Our scheme has brought a tremendous improvement on providing progressive transmission with JPEG-LS. By modifying lossless predictive coding techniques, near-lossless coding based on predictive in conjunction with data hiding, and plus a partially layer scheme.

1 Introduction

LOCO-I (LOw COmplexity LOssless COmpression for Images) is the algorithm at the core of the new ISO/ITU standard for lossless and near-lossless compression of continuous-tone images, JPEG-LS [1].The standard evolved after successive refinements [2, 3, 4, 5, 6], and a complete specification can be found in [7]. Instead of using a simple linear predictor, JPEG-LS use a non-linear predictor that attempts to detect the presence of edges passing through the current pixel and accordingly adjusts prediction. JPEG-LS use a very simple and effective predictor, The Median Edge Detection (MED) predictor that adapts in presence of local edges. This results in a significant improvement in performance in the prediction step.

The latest JPEG efforts in new international standards for lossless and near lossless image compression is represented by JPEG-LS in which the main compression techniques proposed can be broken into the following components:

a. *A prediction* step, in which a value $\hat{y}_{t+1}$ for the next pixel y_{t+1} based on a finite subset (a causal template) of the available past data $y^t = y^1 y^2 \ldots y^t$ (in the sequel, time indexes reflect a raster-scan order).
b. The determination of a *context* in which y_{t+1} occurs (again, a function of a causal template).
c. A probabilistic model for the *prediction residual (or error)* $\varepsilon_{t+1} = y_{t+1} - \hat{y}_{t+1}$, conditioned on the context of y_{t+1}.

This structure was pioneered by the Sunset algorithm [8].

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 463–468, 2005.

1.1 Prediction

In JPEG-LS, prediction is based upon a simple local texture analysis among the three context pixels, *a, b, and c*, as illustrated in Figure 1.

When an edge is detected among the three pixels, the pixel that is not on the edge will be taken as the predictive value. Otherwise, the predictive value will be a well-balanced value drawn from all three pixels. The entire prediction scheme can be described as follows:

```
if (c>= max(a, b)) P = min(a, b);
else{
if (c <= min(a, b)) P = max(a, b); (3)
else P = a+b-c;
}
```

Where *max(a, b)* and *min(a, b)* stands for the maximum value and the minimum value among the two pixels, *a* and *b*, respectively.

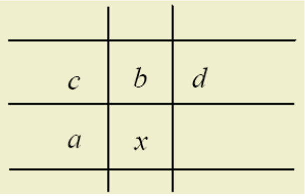


Fig. 1. JPEG-LS predictive template

The prediction consists of a fixed and an adaptive component. The fixed predictor is a median edge detector. i.e. consider b < a, select b as predication of x in many case a vertical edge exists in the image just left of current pixel x.

2 The Proposed Scheme Index-LOCO-I

Here, the proposed near-lossless image data hiding scheme is provided by a remapping the encoding error prior to lossless coding and with high compression ratio and fairly good PSNR. Fig.2 shows the block diagram.

It will first present a preview image after error correction. The pixel after decoding is incorrect due to the prediction error is quantized, and the incorrect reconstruction will also influence to following decoding pixel value and so on. Because of an insufficient predict error will effects the next predict result In order to correct the decoding pixel value, *error correction* step modify the current encoding pixel which will be reconstructed incorrectly. Although we change the original pixel value, the predict condition will be different but quantization is large, the probability of replacement is small so that the predict condition is changed greatly less. If predict error ε quantized q, ε/q will change the reconstruct pixel $\mathbf{I}_{ij}$ to $\mathbf{I}_{ij}/q$, since ε/q won't change the predict pixel P of LOCO-I predict condition $c/q \geq \max(a/q, b/q)$ and $c/q \leq \min(a/q, b/q)$. Consequently, $\varepsilon/q = (\mathbf{I}_{ij}-P_{ij})/q$ will imply $\mathbf{I}_{ij}=-P_{ij}/q+\varepsilon/q=\mathbf{I}_{ij}/q$. While pixels have incorrect reconstruction value through the error correction step, the following pixel will be correct in decoding.

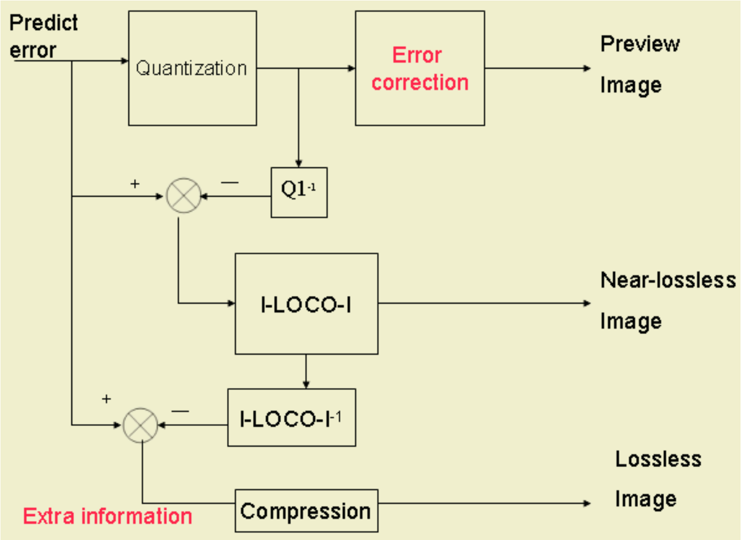


Fig. 2. Block Diagram

2.1 Index LOCO-I (I-LOCO-I) Data Hiding Scheme

We present a data hiding method named Index LOCO-I (I-LOCO-I). In our scheme, original *sign* of prediction error is replaced by *index,* which will be embedded in I-LOCO-I new prediction error *embERR*. An extra information *index_error* is a corrected value for embedding *index* in prediction error values *ERRVAL*. The *index_error* is the difference between the original prediction error and the embedded error. The decoder use *embERR* to reconstruct Near-lossless image without index error.

LOCO-I prediction error can take on any value in the range $-\alpha/2 < \varepsilon < \alpha/2$, where α is the size of the image alphabet. Originally, prediction error via LOCO-I is in the range $-\alpha < \varepsilon < \alpha$, and therefore in average case, I-LOCO-I will embed and reduce the LOCO-I prediction error distribution range to 1/2. Than the entropy coding engine for I-LOCO-I is used Arithmetic coding to compress embedded error.

In a raster scan, image **I**ij after having scanned past data, prediction error $ERRVAL_{ij}$ = **I**ij - Pij will be set index value index following the condition **I**ij >= Pij, index=0, otherwise index=1. Then the algorithm is as follow:

```
for(i=0;i<= image width;i++)
   for(j=0;j<=image height;j++){
       if ( $ERRVAL_{ij}$ is positive )  $index_{ij}$=0;
       else  $index_{ij}$=1 ;
       //Initial the index value for LOCO-I error
       if ( $index_{ij} \oplus 1$ )  N=($ERRVAL_{ij}$-1) / T ;
       else N= ($ERRVAL_{ij}$) / T ;
       $ERRVAL_{ij} = ERRVAL_{ij} - T \times \lceil N/2 \rceil \times (-1)^{N}$;
       $index\_\varepsilon_{ij} = index_{ij} \oplus (N \% 2)$ ;
   }
```

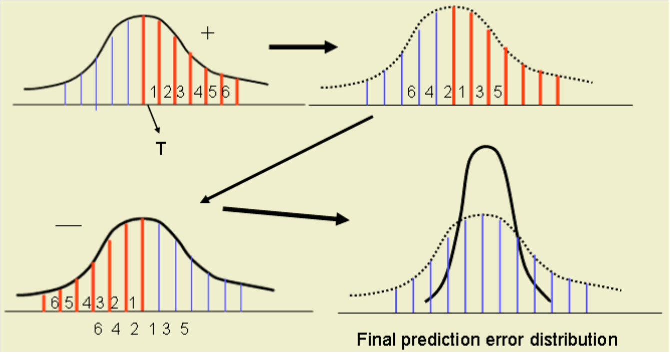


Fig. 3. Index LOCO-I embedding

If prediction error is positive, *embERR* will be an even value, otherwise *embERR* is odd. When T=2, we observe I-LOCO-I presents the best performance. If the image has m×n pixels, LOCO-I will produce m×n index error *index_*ε = (ε % 2) & *index*, which is 0 or 1(here &: means the AND operation). One notable exception to reconstruct image, some pixels is differ to 1, is approximate similar to the original image. Fig.4 shows the entropy of test image and comparison with JPEG-LS.

Unfortunately, the compression of extra information *index_error* is useless. The first-order and second-order entropy H(X), H_2(X) of *index_*ε can be define as

$$H(X) = \sum - index\_\varepsilon(i) \log index\_\varepsilon(i).$$

The above estimates give only a lower -bound on the compression that can be achieved through variable-length coding alone. Differences between higher-order estimates of entropy and the first-order estimate indicate the presence of inter-pixel redundancies.

$$H_2(X) = \sum - index\_\varepsilon(i, j) \log index\_\varepsilon(i, j)$$

Our simulation gives first-order entropy is equal to second-order entropy. H(X)= H_2(X)=0.993. This implies that the entropy of index error is convergence and it is statistic independent (or inter-pixel redundancy is not in the existence of index error).

We would like to stress that the class/style files and the template should not be manipulated and that the guidelines regarding font sizes and format should be adhered to. This is to ensure that the end product is as homogeneous as possible.

3 Simulation Result and Performance Analysis

We take HP's LOCO-I/JPEG-LS implementation V.2.20 of JPEG-LS Software Simulation as our LOCO-I codec. Version 2 supports the lossless and near lossless mode of JPEG-LS. Version 2 also supports color and gray level images. For this experiment, we don't simulation the JPEG-LS "run mode" and calculate the entropy of predict error in place of arithmetic coding.

In Fig.4, I-LOCO-I present a *preview* image plus two layered near-lossless and lossless compression. Since the prediction error is compressed by I-LOCO-I, the bitrate of preview image is greatly low and PSNR is good. And in Fig.5, we can see the progressive three-layer I-LOCO-I has better compression ratio but the bitrate of first layer (preview image) is too large to suit the low-bitrate transmission.

512 ×512	Layer 1	PSNR	Layer 2	PSNR	JPEG-LS
lena	0.13	30.22	3.78	51.61	4.54
woman512	0.30	27.39	3.39	51.77	4.89
baboon	0.86	21.5	4.65	51.27	6.42
barbara	0.47	25.20	3.89	51.52	5.34
airplane	0.09	32.92	2.94	51.88	3.99
pepper	0.06	33.62	3.03	51.73	4.13
2048×2560	Layer 1	PSNR	Layer 2	PSNR	JPEG-LS
bike	0.45	32.14	3.53	51.67	4.81
café	0.90	36.99	4.02	51.56	5.52
woman	0.45	24.39	3.61	51.64	4.88

Fig. 4. The entropy of test image for I-LOCO-I comparison with JPEG-LS

512 ×512	Layer 1	PSNR	Layer 2	PSNR	Layer 3	JPEG-LS
lena	1.04	34.81	1.74	51.61	3.78	4.54
woman512	1.21	33.37	1.95	51.77	4.22	4.89
baboon	1.68	30.07	2.75	51.27	5.49	6.42
barbara	1.36	32.51	2.17	51.52	4.54	5.34
airplane	0.98	36.81	1.47	51.88	3.45	3.99
pepper	1.01	35.94	1.48	51.73	3.48	4.13
2048×2560	Layer	PSNR	Layer 2	PSNR	Layer 3	JPEG-LS
bike	1.26	34.06	1.85	51.67	4.12	4.81
café	1.48	32.24	2.25	51.56	4.73	5.52
woman	1.23	33.58	1.94	51.64	4.17	4.88

Fig. 5. The entropy of test image for our propose comparison with JPEG-LS

4 Conclusion

In this paper lossless or near-lossless compression data hiding techniques based on the criterion of maximum allowable deviation of pixel values are investigate. It is shown how predictive and quantization methods and their combinations can be adaptive to meet the specification of maximum allowable deviation. We have examined lossless compression by modifying lossless predictive coding techniques, near-lossless coding based on predictive in conjunction with data hiding, and a partially layer scheme of lossy plus lossless coding. The predictive approach clearly performs the best. The partially three-layer scheme of lossy plus lossless coding offers a compromise in providing a preview image is perform close to the purely predictive approach.

References

1. J M. J. Weinberger, G. Seroussi, and G. Sapiro, “LOCO-I: A low complexity, context-based, lossless image compression algorithm,” in Proc.1996 Data Compression Conf., Snowbird, UT, Mar. 1996, pp. 140–149.
2. Proposed Modification of LOCO-I for Its Improvement of the Performance, Feb. 1996. ISO/IEC JTC1/SC29/WG1 Doc. N297.

3. Fine-Tuning the Baseline, June 1996. ISO/IEC JTC1/SC29/WG1 Doc.N341.
4. Effects of Resets and Number of Contexts on the Baseline, June 1996.ISO/IEC JTC1/SC29/WG1 Doc. N386.
5. Palettes and Sample Mapping in JPEG-LS, Nov. 1996. ISO/IECJTC1/SC29/WG1 Doc. N412.
6. JPEG-LS with Limited-Length Code Words, July 1997. ISO/IECJTC1/SC29/WG1 Doc. N538.
7. Information Technology—Lossless and Near-Lossless Compression ofContinuous-Tone Still Images, 1999. ISO/IEC 14495-1, ITU Recommend.T.87.
8. J. Villasenor, B. Belzer, J. Liao, "Wavelet Filter Evaluation for Image Compression," IEEE Transactions on Image Processing, Vol. 2, pp. 1053-1060, August 1995.

Feature-Constrained Texturing System for 3D Models

Tong-Yee Lee and Shaur-Uei Yan

Department of Computer Science and Information Engineering,
National Cheng-Kung University, Tainan, Taiwan, Republic of China
tonylee@mail.ncku.edu.tw

Abstract. Significant number of parameterization methods has been proposed to perform good quality of texturing 3D models. However, most methods are hard to be extended for handling the texture mapping with constraints. In this paper, we develop a new algorithm to achieve the matching of the features between the model and texture image. To minimize the distortion artifacts from the matching algorithm, a L2 stretch metric is also applied to optimize the u,v map defined in parameterization domain.

1 Introduction

In computer graphics, texture mapping is a very common technique to enhance the visualization of 3D meshes. Texture mapping adds the detail of pictures to the 3D meshes and let the 3D meshes looks more vividly. Surface parameterization is a common solution to the texture mapping problem. Parameterization commonly maps 3D meshes to a 2D domain that defines the (u, v) texture coordinates in parameterization domain. There is no isometric parameterization to map a general surface patch to a plane [1] and many previous methods [2–10, 16, 21] have been proposed to minimize various kind of distortion for achieving an acceptable visual effect. Most of them do not take feature matching into account. They only provide general solutions to texture mapping without constraints. However, without the correspondence of features, a 3D model with texture coordinates would look strange as shown in Figure 1. For example, the eyes on both texture and model do not match. Therefore, we need to find a method to deal with this constraint issue.

2 Related Work

Many papers have addressed issues in surface parameterization. Surface parameterization can be used for texture mapping. Tutte [15] introduces the barycentric map which guarantees the existence of a one-to-one mapping, i.e., foldover free, for parameterizing a model into the u,v domain. Eck et al. [4] minimize the harmonic energy to approximate the harmonic map and this method can be solved efficiently by a linear sparse matrix. Floater [16] develops the mean value

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 469–475, 2005.

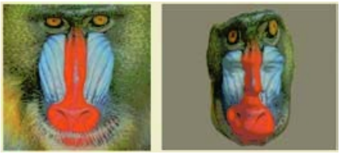


Fig. 1. Texture mapping using a naive non-constrained surface parameterization

coordinates to mimic the discrete harmonic map and this mapping is bijective. Hormann and Greiner [6] derive a formula that measures the distortion of the parameterization and is also invariant for affine transformation. This method does not require the boundary vertices to be fixed on the convex 2D polygon. However, the optimization of solution is slow due to the non-linear property. Sander et al. [3] define a geometric stretch metric for surface parameterization. This approach uses a relaxation approach similar to MIPS [6] to iteratively flatten 3D surfaces. Some papers address the feature correspondence problem between model and texture coordinates. Such topic is called constrained texture mapping. Levy [17] minimizes the distortion and matches features in a least square sense. This approach doesn't exactly match the constrained positions of features and may produce folded triangles in texture domain. Eckstein et al. [18] uses the graph theorem to meet the position constraints, but the method seems too complicated to handle any general case. Kraevoy et al. [19] describes a matching algorithm to align the position of features. This method uses a brute-force approach to satisfy the constraints.

3 Parameterization with Feature Points

In this paper, we develop a novel approach to texture mapping with constraints. Initially, a user only needs to specify several feature correspondences between the model and the texture and then the algorithm would automatically do anything without user-interaction.

3.1 Initial Parameterization

We concentrate the feature-correspondence problem in this paper. The topology of the input 3D mesh is homeomorphic to a disk. If objects are not belonging to this type, it is easy to use extra cutting to form an open disk-like surface, i.e., surface with a boundary [10]. Then, we can use any parameterization method mentioned in the related work. At the current implementation, we adopt the mean value coordinates [16] with a convex boundary in uv domain (see Figure 2 and Figure 3 (c)). Figure 2 shows the weights of the mean value coordinates for each vertex of the model. If the feature points are specified on the border of models, a virtual boundary (Figure 3(d)) can be added to the outside of the parameterized surface boundary to keep all features in the interior of the u,v map. This arrangement can facilitate our method described in Section 3.2.

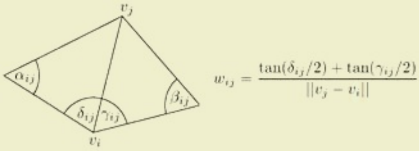

v_j
α_{ij}
δ_{ij}
γ_{ij}
β_{ij}
v_i
$w_{ij} = \frac{\tan(\delta_{ij}/2) + \tan(\gamma_{ij}/2)}{||v_j - v_i||}$

Fig. 2. v_j is the one-ring neighbor of v_i for triangulation mesh

3.2 Feature Matching Using a Partitioning Approach

This section is the core of this paper. The matching idea is mainly based on the procedure of the recursive partitioning in uv domain. In our setup, all computation of the matching stage is performed between the uv map and the texture. After initial parameterization, each vertex of the 3D model has an initial (u, v) coordinate. But the feature vertices don't be matched with corresponding positions on the texture. We wrap the feature vertices to their desired positions by the partitioning approach.

Partitioning. For each partitioning step, we can group the feature vertices into two disjoint sets as follows. First, feature vertices are sorted by the u (or v) coordinate on the texture domain and then the median u_m (or v_m) is computed. Let the line segment L_t: $x = u_m$ (or $y = v_m$) be the partitioning line on the texture domain to separate the feature vertices into two groups G1 and G2 (see Figure 4 (a)). Then we perform the path finding algorithm in the next section to find a path P to isolate the counterparts of G1 and G2 in the uv domain (Figure 4 (b)). A path P has the same starting and ending positions as those of the line L_t. Then, after partitioning by a path P, we can have two disjoint sub-patches SP1 and SP2 in the uv domain. In the next step, we align the path P with the line L_t, i.e., straightening the path P, and then re-parameterize the sub-patches SP1 and SP2 (see Figure 4 (c)). A partitioning step ends. We will iteratively above partitioning tasks on the uv map until each patch contains only one feature point, say f. Finally, we will align each feature f in uv domain with its counterpart f' in the texture domain by a quad-partitioning of each patch as shown in Figure 5. For this purpose, we move a vertex f to a new position f', i.e.,

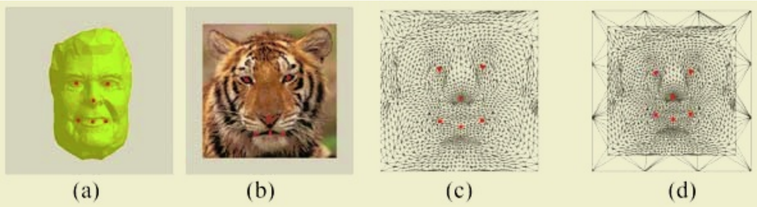


Fig. 3. (a) and (b): feature correspondence between a 3D model and a texture image; (c) and (d): the uv map of the mean value coordinates with and without virtual boundary

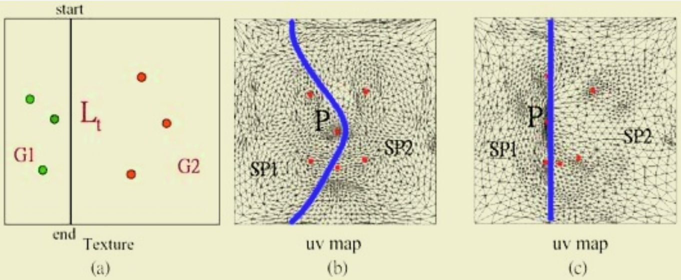


Fig. 4. Each partitioning step determines two separate groups of features in texture and uv domains

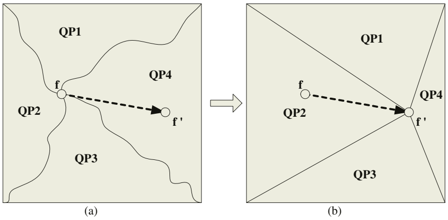


Fig. 5.

correct corresponding position in texture domain. Then, we straighten all paths connecting f and four corners and re-parameterize these four sub-patches.

Finding a partitioning path in uv domain. First, we apply a constrained Delaunay triangulation to the u,v map considering the feature vertices, S, and E (Figure 6 (b)). Both S and E are the starting and ending points for a partitioning path. Second, we compute the minimal spanning trees (MST) for the group G1 and G2 (Figure 6 (c)), respectively. If a MST does not exist, we will apply 1-to-4 subdivision on this triangulation map to have additional paths for finding MSTs. After finding MSTs, we find all edges, i.e., marked by X, in which one of the end points belongs to G1 and the other belongs to G2. When connecting the middle points of these edges, we can form a connected path P like Figure 6 (d). Finally, the starting and ending points of P connect to S and E by the shortest paths L1 and L2 to determine the final partitioning path. This partitioning path divides features into two disjoint groups in uv domain.

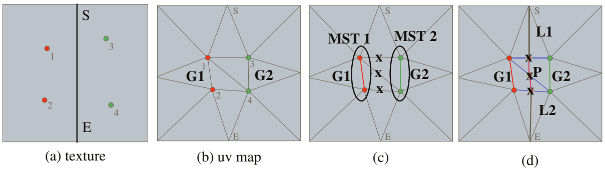


Fig. 6. Finding a partitioning path in u,v domain

3.3 UV Optimization

After matching features, the uv map is distorted to some extent. The uv optimization is a process to improve the uv map as the smoothing stage in [19]. It is an iterative version of parameterization. It moves each vertex except for features within the range of one ring at each iteration to gradually minimize the distortion of uv map. We use the L2 stretch metric [3] instead of the harmonic map in [19]. In our experiment, the L2 stretch metric generally produces better results than the harmonic map (Figure 7).

4 Preliminary Results

We demonstrate some preliminary results using the proposed method. Figure 7 gives the comparison between harmonic map and L2 stretch metric for the smoothing the u,v map. In Figure 7, (c) yields better visual effect than (b). In Figure 8, the positions of corresponding features for an old man and a bear image are very different. The proposed method produces a not bad result. Note that in this example, we specify the features on the border of texture image, therefore we need to add virtual boundary for the u,v map. Finally, we show another interesting example in Figure 9. In this case, we do not require virtual boundary. We perform the experiments on a PC with Pen-tium IV 2.4 GHz and 512 MB RAM. On the average, it takes a minute to finish a texture mapping for these examples.

5 Conclusion

In this paper, we have presented a new algorithm for the constrained texture mapping. Preliminary results show that this new method is very promising. We

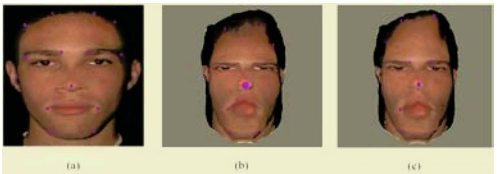


Fig. 7. Smoothing u,v map. (a):original texture, (b):smoothed with a harmonic map and (c):smoothed with L2 stretch metric

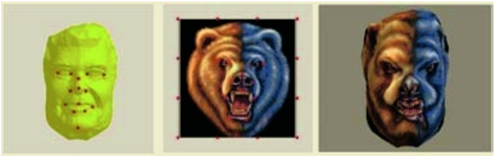


Fig. 8. Texturing an old man with a bear image

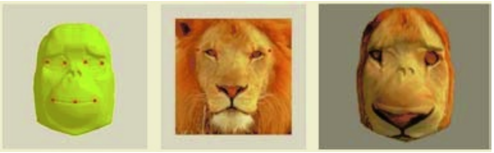


Fig. 9. Texturing a monkey face with a lion image

can successfully handle texture mapping with constraints well. In future, there are many works to be done. For example, there is also a need for matching correspondence in 3D metamorphosis application and consistent surface parameterization. We will plan to explore the possibility of our approach to these important applications. Furthermore, in the current method, we need many re-parameterizations of the patches and therefore the computation cost can be expensive as the features are increased drastically. We would like to find a better approach to reduce the number of re-parameterization in near future. Another interesting and our ongoing research is to compute progressive texture transfer between two models in metamorphosis applications [22, 23].

Acknowledgement

This paper is supported by the National Science Council, Taiwan, Republic of China, under contract No. NSC-93-2213-E-006-026 and NSC-93-2213-E-006-060.

References

1. L. V. Ahlfors and L. Sario; *Riemann Surfaces*; Princeton University Press, Princeton, New Jersey, 1960
2. J. Maillot, H. Yahia, and A. Verroust; *Interactive texture mapping*; Proceedings of SIGGRAPH, 1993, pp. 27-34
3. P. Sander, J. Snyder, S. Gortler and H. Hoppe; *Texture mapping progressive meshes*; Proceedings of SIGGRAPH, 2001, pp. 409-416
4. M. Eck, T. DeRose, T. Duchamp, H. Hoppe, M. Lounsbery, and W. Stuetzle; *Multiresolution analysis of arbitrary meshes*; Proceedings of SIGGRAPH, 1995, pp. 173-182
5. M. S. Floater; *Parametrization and smooth approximation of surface triangulations*; Computer Aided Geometric Design, 14(3):231-250, 1997
6. K. Hormann and G. Greiner; *Mips: an efficient global parameterization method*; Curve and Surface Design: St. Malo 1999, pages 153-162, Vanderbilt University Press, 2000
7. Sorkine, D. Cohen-Or, R. Goldenthal, and D. Lischinski; *Bounded-distortion Piecewise Mesh Parameterization*; IEEE Visualization, 2002, pp. 355-362
8. G. Zigelman, R. Kimmel, and N. Kiryati; *Texture mapping using surface flattening via multidimensional scaling*; IEEE Transactions on Visualization and Computer Graphics, Vol. 8, No. 2, pp. 198-207, 2002

9. G. Piponi and D. Borshukov; *Seamless Texture Mapping of Subdivision Surfaces by Model Pelting and Texture Blending*; Proceedings of SIGGRAPH, 2000, pp. 471-478
10. X. Gu, S. J. Gortler, and H. Hoppe; *Geometry Images*; Proceedings of SIGGRAPH, 2002, pp. 355-361
11. C. Gotsman, X. Gu, and A. Sheffer; *Fundamentals of Spherical Parameterization for 3D Meshes*; Proceedings of SIGGRAPH, 2003, pp. 358-363
12. T. Y. Lee and P. H. Huang; *Fast and Institutive Polyhedra Morphing Using SMCC Mesh Merging Scheme*; IEEE Transactions on Visualization and Computer Graphics, Vol. 9, No. 1, pp. 85-98, 2003
13. Sheffer and E. D. Sturler; *Smoothing an Overlay Grid to Minimize Linear Distortion in Texture Mapping*; ACM Transactions on Graphics, Vol. 21, Issue 4, pp. 874-890, 2002
14. U. Pinkall and K. Polthier; *Computing discrete minimal surfaces and their conjugates*; Experimental Mathematics, 2(1):15-36, 1993
15. W. Tutte; *Convex representation of graphs*; In Proc. London Math. Soc., volume 10, 1960
16. M. S. Floater; *Mean value coordinates*; Computer Aided Geometric Design, 20(1):19-27, 2003
17. B. Levy; *Constrained Texture Mapping for Polygon Meshes*; ACM SIGGRAPG 2001, 417-424
18. I. Eckstein, V. Surazhsky, and C. Gotsman; *Texture Mapping with Hard Constraints*; Computer Graphics Forum 20, 3, 95-104
19. V. Kraevoy, A. Sheffer, and C. Gotsman; *Matchmaker: constructing constrained texture maps*; ACM SIGGRAPH 2003, 326-333
20. J. Pach and R. Wenger; *Embedding planar graphs with fixed vertex locations*; Proceedings of Graph drawing '98. Lecture Notes in Computer Science 1547, Springer-Verlag, 1998, 263-274
21. Tong-Yee Lee and Shaur-Uei Yan; *Texture Mapping on Arbitrary 3D Surfaces*; Lecture Notes on Computer Science 3024, Springer-Verlag, pp. 721-730, 2004
22. Tong-Yee Lee and P.H Huang; *Fast and Institutive Polyhedra Morphing Using SMCC Mesh Merging Scheme*; IEEE Transactions on Visualization and Computer Graphics, Vol. 9, No. 1, pp. 85-98, 2003
23. Chao-Hung Lin and Tong-Yee Lee; *Metamorphosis of 3D Polyhedral Models Using Progressive Connectivity Transformations* ; IEEE Transactions on Visualization and Computer Graphics, Jan./Feb. Issue, Vol. 11, No.1, pp. 2-12, 2005

Dynamic Integrated Model for Distributed Multimedia System

Ya-Rong Hou and Zhang Xiong

601 Box, Beijing University of Aeronautics and Astronautics
Haidian District, Beijing 100083, P.R. China
yaronghou@sohu.com

Abstract. Multimedia information and communication network are two key elements of a distributed multimedia system. This paper researches distributed multimedia system from the points of view of both multimedia data and network channels. A dynamic model *SCUDM* for distributed multimedia system is proposed and a scheduling algorithm *MCD* is put forward. Simulation experiment shows that algorithm *MCD* has an advantage over algorithm *EDF* on the network utilization.

1 Introduction

So far many kinds of distributed multimedia systems (DMS) have been widely used, such as video/audio on demand system, video conference system, telemedicine system, etc. But because of the limitation of network bandwidth, packet loss, and delay and jitter of data transmission over network, it's hard to provide high quality of service. To overcome this problem, researches are mainly focused on source coding [1-3], channel coding [2-4], and quality of service controlling [5, 6], where multimedia data and network channel are often researched separately. In fact, there are close relationships between multimedia data and network channel, because in a DMS it is the network channel that transmits multimedia data from source to destination. Both of these two elements affect the performance of a DMS.

Therefore, the major objective of this paper is researching DMS from the points of view of both multimedia data and network channel. Section 2 introduces the conception of synchronization interval unit and the process of multimedia communication and presentation of a DMS. In section 3, a dynamic model for DMS is proposed. In section 4, a scheduling algorithm is put forward. Section 5 is performance evaluation of the algorithm. Section 6 concludes this paper.

2 Description of Distributed Multimedia System

Different terminal nodes of a DMS are connected with each other by communication network. Different kinds of data are transmitted among these terminals. Although there are many kinds of data, such as text, image, video, audio, etc., they can be divided into two categories, anisochronous (discrete) data (e.g. text and images) or

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 476–482, 2005.

isochronous (continuous) data (e.g. video and audio). Isochronous data has strict temporal restrictions. Inter-stream synchronization and intra-stream synchronization describe these temporal restrictions. There are some common properties among these two categories of data. First of all, they all have play-out deadline and presentation duration. Secondly, the communication and presentation process are similar.

2.1 Abstract Description of Multimedia Information

Discrete data and continuous data can be decomposed into smaller units. This smaller unit is named synchronization interval unit (SIU) [7, 8]. SIU is the basic data unit for multimedia presentation. It is clear that every SIU has some fixed properties, such as size, play-out deadline, play-out duration, etc. A video frame can be selected as the video SIU. A sample or a series of audio samples can be taken as the audio SIU. As for discrete data, the whole object can be seemed as the SIU.

2.2 Communication and Presentation Process of a DMS

Multimedia data, i.e. SIU, is transmitted over communication subnet and arrives at the destination. SIUs will be played back with fixed playback schedule. Because of the random delays of network transmission, SIUs can not always arrive at the scheduled time. Every destination node has buffer with fixed size. If a SIU arrives before its playback time and the destination buffer has space to store it, the SIU will enter the buffer waiting for playback. If the SIU arrives behind its presentation deadline, it will be discarded directly. Fig. 1 describes the communication and presentation process of multimedia data.

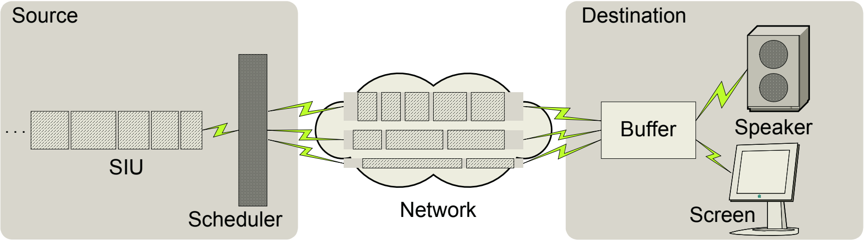


Fig. 1. Communication and Presentation Process of Multimedia data within a DMS

3 Dynamic Integrated Model for Distributed Multimedia System

In a DMS, multimedia information is transmitted over communication network. Multimedia information and communication network are two basic elements of a DMS, so a DMS can be abstractly expressed with these two elements.

3.1 Model of Communication Network

Communication network is a very important factor of a DMS. Every channel has some attributes, among which bandwidth and propagate delay are the most important two attributes. A network with QoS assurance has fixed number of channels and every channel has fixed bandwidth and fixed propagate delay. Communication network can be expressed as follows:

$$\begin{aligned} &N = \{C_i \mid 1 \le i \le n\} \\ &C_i =< c_i, \varepsilon_i >, 1 \le i \le n \end{aligned} \tag{1}$$

Formula (1) means network *N* has *n* channels and the *i*th channel C_i can be expressed with bandwidth c_i and propagate delay ε_i.

3.2 Model of Multimedia Data

SIU is the basic unit of multimedia information, so multimedia information *M* can be described as a set of SIUs. Every SIU has some properties, among which SIU size, presentation deadline, transmission time and arrival time are key attributes that directly affect channel selecting and SIU sending time during SIU scheduling procedure. Multimedia information can be expressed as follows:

$$\begin{aligned} &M = \{U_j \mid 1 \le j \le m\} \\ &U_j =< s_j, p_j, t_j, a_j >, 1 \le j \le m \end{aligned} \tag{2}$$

Here U_j is the *j*th SIU and U_j can be described with SIU size s_j, presentation deadline p_j, transmission time t_j and arrival time a_j. If U_j is transmitted over channel C_i at time t_j, the arrival time a_j can be worked out using equation: $a_j = t_j + {}^{s_j}\!/_{c_i} + \varepsilon_i$.

3.3 Scheduling of Multimedia Data over Communication Network

During the communication procedure, different channel transmits different SIUs. When transmission procedure finished, SIUs which were transmitted over the same channel formed a SIU scheduling sequence. So there are *n* corresponding SIU queues and a SIU scheduling queue is a list of SIUs.

$$\begin{aligned} &R : C_i \to Q_i, 1 \le i \le n \\ &Q_k \cap Q_l = \Phi, 1 \le k \le n, 1 \le l \le n, k \ne l \\ &\bigcup_{i=1}^{n} Q_i = \{U_j \mid 1 \le j \le m\} \end{aligned} \tag{3}$$

Here, *R* represents the one-to-one mapping between channel C_i and the SIU scheduling queue Q_i.

In initial state, all SIUs are in source station, i.e. $Q_i = \Phi, 1 \le i \le n$. When the transmission procedure begins, U_j is sent out over an idle channel C_i, and queue Q_i relevant to C_i is expanded, this process can be expressed with $Q_i = Q_i \cup \{U_j\}$. When the communication procedure concludes, the expanding process of Q_i will conclude too. For every $k, 1 \le k \le n$, every $l, 1 \le l \le n$, when $k \ne l$, $Q_k \cap Q_l = \Phi$, and $\bigcup_{i=1}^{n} Q_i = \{U_j \mid 1 \le j \le m\}$, i.e. $\bigcup_{i=1}^{n} Q_i = M$.

3.4 SIU-Channel-United Dynamic Model

Finally, we get the dynamic model for DMS, namely SIU-Channel-United Dynamic Model (*SCUDM*).

$$\begin{aligned} & SCUDM = \langle M, N, R \rangle \\ & M = \{U_j \mid 1 \le j \le m\} \\ & N = \{C_i \mid 1 \le i \le n\} \\ & R : C_i \to Q_i, 1 \le i \le n \\ & U_j =< s_j, p_j, t_j, a_j >, 1 \le j \le m \\ & C_i =< c_i, \varepsilon_i >, 1 \le i \le n \\ & Q_k \bigcap Q_l = \Phi, 1 \le k \le n, 1 \le l \le n, k \ne l \\ & \bigcup_{i=1}^{n} Q_i = \{U_j \mid 1 \le j \le m\} \end{aligned} \tag{4}$$

Model *SCUDM* describes not only the static attributes of a distributed multimedia system, but also the dynamic SIU scheduling process by constructing the one-to-one mapping relationships between the network channels and the SIU scheduling queues.

Model *SCUDM* has some characteristics. First of all, it is media-type-independent. SIU is the basic unit of multimedia data. Multimedia streams can be considered as a series of SIUs. To single media stream, different SIUs have different presentation deadlines. To concurrent multimedia streams, different SIUs may have same presentation deadlines. All kinds of multimedia data can be uniformed as SIUs. Secondly, it is coding-algorithm-independent. No matter what kind of coding criterion (MPEG-x or H.26x) one multimedia application uses, the basic unit that is meaningful for playback device and for presentation process is SIU. Thirdly, it is independent of network topological structure. A group of channels are used to represent the network, taking no account of the topology of the network.

4 Multimedia Data Scheduling Algorithms

Based on model *SCUDM*, the process of multimedia data transmission over network can be clearly described. A multimedia data scheduling algorithm *MCD* (Minimal Cluster Delay Scheduling Algorithm) is proposed.

a) Initializing the SIUs queue of source node. At the beginning, all SIUs locate in source node, so $M_0 = M$. Here, M_0 is the set of SIUs that are not scheduled by the source scheduler. Sort SIUs in nondecreasing order of their playback deadlines. That is, $\forall k, \forall l, 1 \le k < l \le m$, $p_k \le p_l$.

b) Initializing the set of idle channels. At the beginning, all channels are idle, so $N_0 = N$. Here, N_0 is the set of idle channels.

c) Initializing the scheduling queue of SIUs. In initial state, all SIUs are in source station, i.e. $Q_i = \Phi, 1 \le i \le n$.

d) Selecting of SIUs that will be scheduled. When $M_0 \ne \Phi \wedge N_0 \ne \Phi$, let $n' = \min(|M_0|, |N_0|)$, schedule the first n' SIU of M_0, and update set M_0 correspondingly.

e) Selecting of idle channel. Select channel for every SIU that will be scheduled to minimizes the gross end-to-end transmission delay of SIUs, i.e., minimize the value of $\sum_{j'=j}^{j+n'-1}(t_{j'}+\frac{s_{j'}}{c_{i_{j'}}}+\varepsilon_{i_{j'}})$. Update set N_0 and Q_i correspondingly . As soon as there is an idle channel, the SIU in M_0 will be scheduled immediately.

f) Repeat steps a)~e) until $M_0=\Phi$.

5 Performance Evaluation

In order to evaluate the proposed algorithm *MCD*, a simulation experiment is executed. We compare the performance of algorithm *MCD* with the performance of the traditional scheduling algorithm *EDF* (earliest deadline first).

Table 1 and table 2 are the configurations of SIUs and of network used by simulation experiment. (data in table 1 and in table 2 come from literature [8].)

Table 1. Configurations of SIUs

Configurations	Object	Number of SIUs	Mean size of SIUs (kbyte)	Range of variation of SIUs (kbyte)
1#	Vedio	2500	1	[0.5, 1.5]
	Audio	2500	0.268	constant
2#	Vedio	2500	2	[1.0, 3.0]
	Audio	2500	0.268	constant
3#	Vedio	2500	4	[2.0, 6.0]
	Audio	2500	0.268	constant

Table 2. Configurations of network

Configurations	Number of channels	Capacity of channels (Mbps)		
		Channel C_1	Channel C_2	Channel C_3
1#	2	0.6	0.6	-
2#	2	1.0	0.6	-
3#	2	1.0	1.0	-
4#	2	1.5	1.0	-
5#	3	0.4	0.4	0.4
6#	3	0.6	0.6	0.4
7#	3	1.0	0.6	0.4
8#	3	1.5	1.0	0.6

Fig. 2 is the comparison of network usage factor. Fig. 3 is the comparison of multimedia data unit presentation ratio, which stands for the ratio of multimedia data units number that are played back on time to the total of multimedia data units.

From Fig. 2 we can get that the network usage factor of algorithm *MCD* is higher than that of algorithm *EDF*. The reason is that *MCD* aims to minimize the gross end-to-end transmission delay of SIUs. Fig. 3 shows that algorithm *EDF* has better performance on multimedia data unit playback ratio but the difference between is not very obvious. The reason is that minimizing the gross end-to-end transmission delay of SIUs shortens the process of data communication, at the same time the destination buffer is more likely to overflow and the data unit playback ratio descends. Besides, unit playback ratio has direct relations with the destination buffer size. In our simulation experiment, the buffer size is relatively small (equal to 5 times average SIU size) and that's why the result of the experiment is not satisfying.

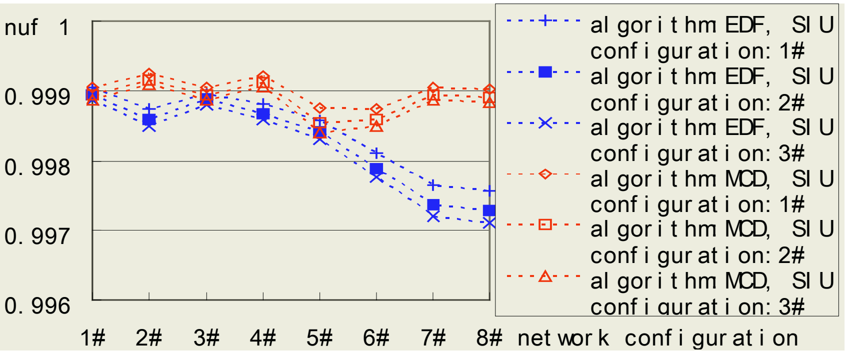


Fig. 2. Comparison of Network Usage Factor

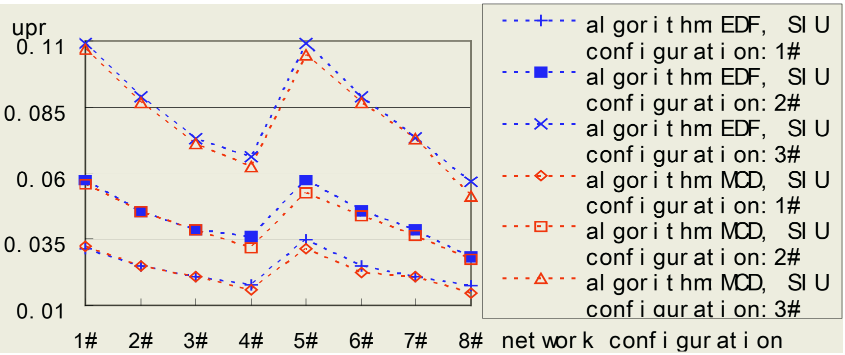


Fig. 3. Comparison of Multimedia Data Unit Playback Ratio

6 Conclusion

Although many kinds of distributed multimedia systems have appeared, the quality of service is not satisfying. In this paper, model *SCUDM* is proposed. Based on the model a scheduling algorithm *MCD* is put forward. Simulation experiment shows that *MCD* has an advantage over *EDF* on the network usage factor.

References

1. B. Fong, G. Y. Hong, A. C. M. Fong, Constrained error propagation for efficient image transmission over noisy channels, *IEEE Transactions on Consumer Electronics*, 48(1), 2002, 49-55
2. 2. T. P. Chen, T. Chen, Adaptive joint source-channel coding using rate shaping, Proc. 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, USA, 2002, 1985-1988
3. K. Takahata, N. Uchida, Y. Shibata, QoS control of multimedia communication over wireless network, Proc. 22nd International Conference on Distributed Computing Systems Workshops, Vienna, Austria, 2002, 336- 340
4. C. Fan, H. Cui, K. Tang, Error-correcting for variable-length codes based on unequal error protection, Electronics Letters, 39(2), 2003, 221- 222
5. M. King, Breaking the server and data communications barrier with serverless guaranteed quality of service (GQoS) compliant communications, Proc. 1st International Conference on Peer-to-Peer Computing, Linkoping, Sweden, 2001, 36-44
6. Q. Zhang, Y. Q. Zhang, W. Zhu, Resource allocation for audio and video streaming over the Internet, Proc. IEEE International Symposium on Circuits and Systems, Geneva, Switzerland, 2000, 21-24
7. M. Woo, N. U. Qazi, A. Ghafoor, A synchronization framework for communication of pre-orchestrated multimedia information, IEEE Network Mag., 8(1), 1994, 52-61
8. S. Baqai, M. F. Khan, M. Woo, S. Shinkai, A. A. Khokhar, A. Ghafoor, Quality-based evaluation of multimedia synchronization protocols for distributed multimedia information systems, IEEE Journal on Selected Areas in Communications, 14(7), 1996, 1388-1403

Joint Detection for a Bandwidth Efficient Modulation Method

Zhonghui Mei, Lenan Wu, and Shiyuan Zhang

Department of Radio Engineering, Southeast University, Nanjing, Postfach 210096, China
mei_hui@sohu.com

Abstract. A bandwidth efficient communication model is propoded via a narrowband band-pass filter at the transmitting end of the system. In order to reduce the ISI caused by the filter, a joint detection algorithm based on low-density parity check matrix codes (LDPC) is presented.

1 Introduction

Most transmission systems have band limitations imposed by either the nature bandwidth of the transmission medium or by regulatory conditions. Therefore, the challenge in data transmission systems is to obtain the highest possible data rate in the bandwidth allotted with the least number of errors. In this letter, a narrowband band-pass filter (NBPF) is added at the transmitting end of the communication system to achieve a bandwidth efficient transmission and suppress the interference to other channels, as is illustrated in Fig 1. To compensate for the ISI resulted from the filter, we introduce a joint detection algorithm by modifying the Sum-product algorithm (SPA).

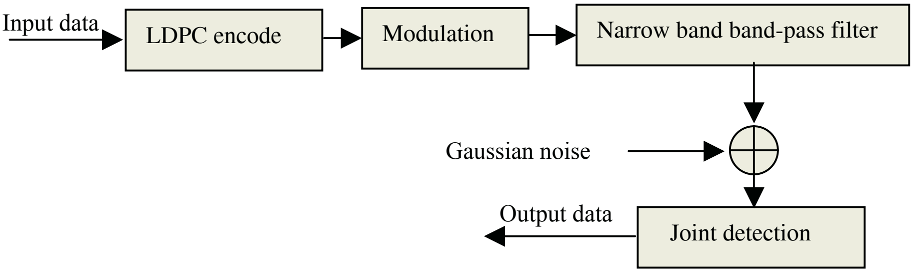


Fig. 1. The bandwidth efficient communication model

2 Signal Pulse

The signal pulse proposed here (in the time domain) is defined as

$$g(t) = \begin{cases} 0 & , 0 \le t \le T/4 \\ g_{mid}(t) & , T/4 \le t \le 3T/4 \\ 0 & , 3T/4 \le t \le T \end{cases} \quad (1)$$

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 483–487, 2005.

where $g_{mid}(t)=0.5\left[1+\cos\left(\frac{4\pi}{T}(t-T/2)\right)\right]$, is the raised cosine pulse [1]. The power spectrum density (PSD) of the modulated signals with pulse g(t) is plotted in Fig 2. When this signal is filtered by the NBPF, its corresponding spectrum is illustrated in Fig 3.

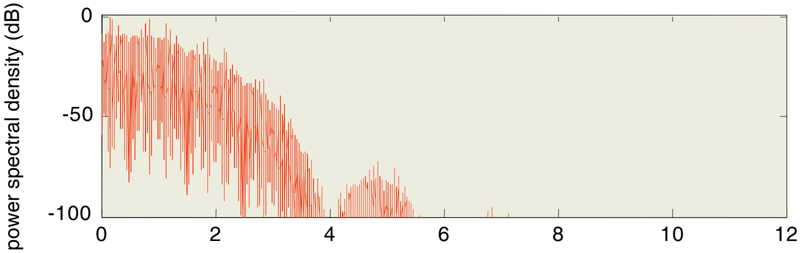


Fig. 2. PSD of the modulated signals with pulse g(t)

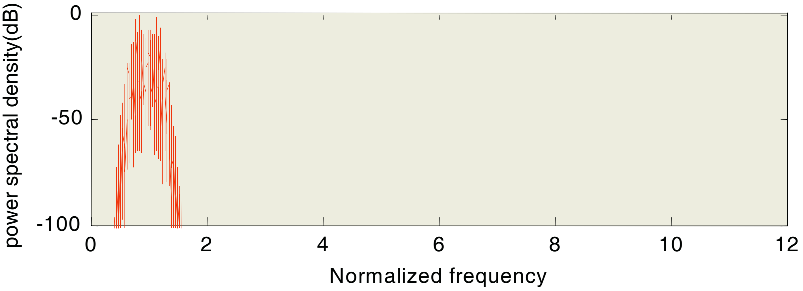


Fig. 3. PSD of the filtered signals

Assuming the linear phase NBPF has LK taps (K denotes the number of samples per signal pulse), thus LK/2 delay is introduced by the filter. The output of the filter which contains the input information b(i) can be expressed as

$$\begin{aligned} f(iK+LK/2-K/2) &= \sum_{l=0}^{L-1}\sum_{k=0}^{K-1} b(i+\frac{L-1}{2}-l)g(k) \\ &\quad\times h(LK-lK-k) \\ &= b(i)\sum_{\boldsymbol{k}=0}^{K-1} g(\boldsymbol{k})h(LK/2+K/2-\boldsymbol{k}) \\ &\quad+\sum_{\substack{l=0\\ l\neq(L-1)/2}}^{L-1}\ \sum_{k=0}^{K-1} b(i+\frac{L-1}{2}-l)g(k)h(LK-lK-k) \end{aligned} \tag{2}$$

where $b(i)\in\{-1,1\}$. $g(k)$ denotes the samples of g(t) at times t=kT/K where k=0,1,…,K-1. {h(j)} is the pulse response of the filter.

The first term on the left side of (2) represents the desired signal; and the second term represents the interference resulting from the neighbouring bits.

3 Joint Detection Algorithm

To suppress the ISI, a joint detection method is derived by modifying the SPA algorithm in [2], which is illustrated based on the Tanner graph [3].

The samples of the receiving signals are given by

$$r(i) = f(iK + LK/2 - K/2) + n(i) \qquad i = \frac{L\text{-}1}{2}, \frac{L\text{-}1}{2} + 1, \ldots \tag{3}$$

where n(i) is the additive Gaussian noise with variance σ^2.

In order to derive $p(b(i) = x \mid r(i))$, firstly we define

$$\begin{aligned}\mathfrak{R}(b(i) = x, \bar{\hat{b}}_{/i}) &= x\sum_{k=0}^{K\text{-}1} g(k)h(LK/2 + K/2 - k) \\ &+ \sum_{\substack{l=0 \\ l\neq(L\text{-}1)/2}}^{L\text{-}1} \sum_{k=0}^{K\text{-}1} \hat{b}(i + \frac{L-1}{2} - l)g(k)h(LK - lK - k)\end{aligned} \tag{4}$$

where $\bar{\hat{b}}_{/i} = \left[\hat{b}(i - \frac{L\text{-}1}{2}), \ldots, \hat{b}(i\text{-}1), \hat{b}(i+1), \ldots, \hat{b}(i + \frac{L\text{-}1}{2})\right]$ and $\hat{b}(i) \in \{-1,+1\}$.

So that

$$\begin{aligned}p(b(i) = x \mid r(i)) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \times \\ &\sum_{all\ \bar{\hat{b}}_{/i}} \left\{ \exp\left(-\frac{(r(i) - \mathfrak{R}(b(i) = x, \bar{\hat{b}}_{/i}))^2}{2\sigma^2}\right) \prod_{\hat{b}(l)\in\bar{\hat{b}}_{/i}} p(\hat{b}(l)) \right\}\end{aligned} \tag{5}$$

According to the sum-produce algorithm, the message passing from variable node to check node can be expressed as

$$\begin{aligned} q_{ij}^{iter}(x) &\propto p(b(i) = x \mid r(i)) \times \prod_{j'\in C_{i\backslash j}} t_{j'i}^{iter}(x) \\ &\propto \sum_{all\ \bar{\hat{b}}_{/i}} \left\{ \exp\left(-\frac{(r(i) - \mathfrak{R}(b(i) = x, \bar{\hat{b}}_{/i}))^2}{2\sigma^2}\right) \prod_{\hat{b}(l)\in\bar{\hat{b}}_{/i}} p(\hat{b}(l)) \right\} \times \left(\prod_{j'\in C_{i\backslash j}} t_{j'i}^{iter}(x) \right) \\ &\propto k_{ij} \sum_{all\ \bar{\hat{b}}_{/i}} \left\{ \exp\left(-\frac{(r(i) - \mathfrak{R}(b(i) = x, \bar{\hat{b}}_{/i}))^2}{2\sigma^2}\right) \prod_{\hat{b}(l)\in\bar{\hat{b}}_{/i}} Q^{iter-1}(\hat{b}(l)) \right\} \\ &\times \left(\prod_{j'\in C_{i\backslash j}} t_{j'i}^{iter}(x) \right) \end{aligned} \tag{6}$$

where k_{ij} is the normalized constant to ensure that $q_{ij}^{iter}(1)+q_{ij}^{iter}(-1)=1$. $C_{i\backslash j}$ represents the set of check nodes connected to variable node i other than check node j. $t_{j'i}^{iter}(x)$ denotes the information passing from check node to variable node and is given by (7). $Q^{iter-1}(\hat{b}(m))$ is the outgoing information from the last iteration and is given by (8).

As is the same with the SPA algorithm, the message passing from check node to variable node $t_{ji}^{iter}(x)$ and the outgoing message $Q^{iter}(b(i)=x)$ are expressed as [2]

$$t_{ji}^{iter}(x)=1/2+1/2\prod_{i'\in V_{j\backslash i}}(1-2q_{i'j}^{iter-1}(x)) \tag{7}$$

where $V_{j\backslash i}$ denotes the variable nodes connected to check node j other than variable node i.

$$Q^{iter}(b(i)=x)\propto p(b(i)=x\,|\,r(i))\times\prod_{j\in C_i}t_{ji}^{iter}(x) \tag{8}$$

where C_i denotes the check nodes connected to variable node i.

4 Simulation Results

The block length of the LDPC is 4000 with code rate 1/2. The bandwidth of the linear phase NBPF is designed to be [15/(16T), 17/(16T)].

The performance comparison of different detection methods are presented in Fig 4. It demonstrates that significant performance gains can be obtained by the joint detection algorithm. This effect can be explained by noting that the joint detection algorithm takes the information fed back from the LDPC decoder to eliminate the ISI.

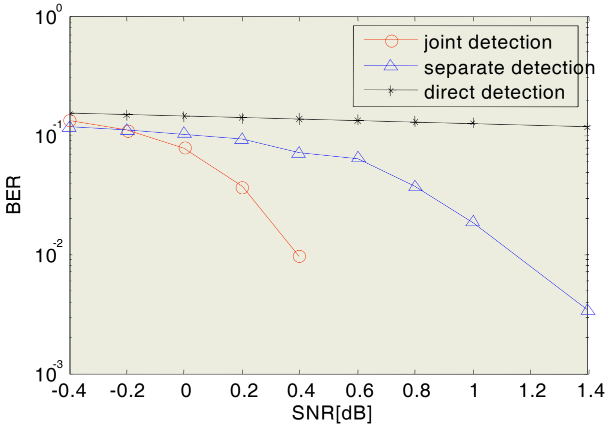


Fig. 4. Performance of different detection algorithms

5 Conclusions

In this letter, we provide a bandwidth efficient communication model. A joint detection algorithm can compensate for the ISI introduced by the narrowband bandpass filter.

References

1. J. G. Proakis: Digital Communications, 4th ed.New York: McGrwawHill (2000).
2. William E.Ryan : An introduction to LDPC Codes, CRC Press (2004).
3. R. M. Tanner : An Recursive Approach to Low Complexity Codes, IEEE Trans. Inform. Theory. (1982) 533-547.

An Efficient and Divisible Payment Scheme for M-Commerce*

Yong Zhao[1], Zhen Han[1], Jiqiang Liu[1], and Zhigang Li[2]

[1] School of Computer Information and Technology, Beijing Jiaotong University
[2] The School of Telecommunication Engineering, Xidian University

Abstract. Almost all of the mobile devices have some fixed characters which can be distinguished easily, and also they are so portable that can be taken with yourself, so the mobile device which is used in electronic business is always the same, as the result, electronic cash, consumer and mobile device can be tied together, and we can check whether the cash belongs to the user by authenticating the fixed character of the mobile device. In this paper, a portable and divisible electronic payment scheme is presented under the idea above, whose computational load, storage needed and network load are all light, so it is fit for mobile commerce. In addition, an augmented dynamic password identity authentication scheme is given after disclosing the hidden security trouble of SDPA. Meanwhile, a new identity-based digital signature scheme is proposed.

1 Introduction

M-Commerce, including electronic payment, electronic contract and so on, is a kind of electronic commerce which is achieved in the way of connecting to Internet with mobile device, such as mobile phone and Personal Digital Assistant (PDA). In recent years, Internet is developed rapidly and mobile device is used more and more wildly, so the potential development of M-commerce is unimaginable. While there is a few disadvantages of mobile device, that is their storage resource and computational resource is limited and the bandwidth of wireless transportation is low, so that a lot of protocols, which run perfectly in fixed terminals, cannot be transplanted into wireless network. As the result, presenting a new protocol suitable for wireless network is necessary.

As we all know, almost all of the mobile devices have some fixed characters which can be distinguished easily and also they are so portable that are easier to be carried, as a result the mobile device which is used in electronic business hardly changes, so electronic coin, consumer and mobile device can be tied together, and we can judge whether the coin belongs to the user by checking the fixed character of the mobile device. In this paper, a portable and divisible electronic payment scheme is presented based on the idea above, at the same time, we also present a more secure dynamic password authentication scheme and a new identity based digital signature algorithm.

* Supported by The State Key Lab of Modern Communication (51436040203DZ01) and The "973" National Fundamental Research Program of China (TG1999035801)

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 488–496, 2005.

During the course of the payment, including withdraw, paying, deposit, only two identity based digital signature signing and verifying computation are needed, so compared with the scheme in [1,2,3], this scheme has great advantage in computational load, storage needed and network load, obviously, it is suitable for M-commerce. In addition, an efficient electronic payment protocol is given in [4] which is also suitable for wireless e-commerce, while the e-cash in that scheme is indivisible, and also it require the user pay on line, this condition has limited the user's demand.

The rest of this paper is organized as follows. In the second part, a hidden trouble is found after the security analysis of SDPA dynamic password authentication scheme, and then a way to get rid of it is given. In the third part, we have briefly introduced the identity based cryptosystem, while so far, the way always used to implement it is Weil-pairing or Tate-pairing, and the computational load is a little heavy, therefore, a new identity based digital signature scheme suitable for wireless network is proposed. In the forth part, our own electronic payment scheme is presented with the security and efficiency analysis. Finally, I will conclude this paper.

2 Dynamic Password Authentication

It is simple and convenient to authenticate user's identity with password, while the static password authentication scheme we always use is subject to eavesdropping, dictionary attack, replay attack and so on, so it's insecure to be used in Internet. Therefore, a series of identity authentication schemes are proposed, including dynamic password identity authentication, one-time key. And in dynamic password authentication, the password given by the user is different at different time, on different event, after different request; as a result, the attacks mentioned above can be avoided. The implementation of dynamic password authentication scheme can be put into two categories: one is based on cryptographic algorithm, and the other is based on some simple computation, such as Hash function or Xor operation. Obviously, the latter's computational load is light, so it is an perfect choice for mobile device.

In 2000, Sandiringama [5] put forward the first relatively mature and simple dynamic identity authentication scheme SAS, which is based on simple hash function, while it is subject to replay, dictionary, DoS attack [6,7,8]. In 2002, D.Yuan improved the scheme and got SDPA scheme, unfortunately, the improved scheme is also subject to man-in-middle attack, here we'd like to show the scheme and then give the security analysis with respect to it, finally give out our own augmented scheme.

2.1 Secure Dynamic Password Authentication (SDPA)

Register Phrase

1) A user U chooses his (or her) password s and a random number N_0, then computes $H^2(s \parallel N_0)$, and sends his (or her) ID, N_0 and $H^2(s \parallel N_0)$ to identity authenticating server S via a secure channel;

2) The server gets the user's ID, N_0 and $H^2(s \parallel N_0)$ saved, note that $H^2(s \parallel N_0)$ here is used as U 's authenticating initiator.

Identity Authenticating Phrase

1) *U* sends his request to the server;

2) The Server *S* randomly chooses M_n and computes $M_n \oplus H^2(s \parallel N_n)$, then sends N_n and $M_n \oplus H^2(s \parallel N_n)$ to the user;

3) *U* firstly computes $H^2(s \parallel N_n)$, then it is easy for him to get M_n, and then he randomly chooses N_{n+1}, computes: $X = H(s \parallel N_n) \oplus H(M_n)$, $Y = H^2(s \parallel N_{n+1})$ $\oplus H^2(M_n)$, $Z = H(H^2(s \parallel N_{n+1}) \parallel N_{n+1})$, $N = N_{n+1}$, and sends them to the server together.

4) The server *S* calculates $H(X \oplus H(M_n))$ to authenticate the user, computes $Y \oplus H^2(M_n)$ to get $H^2(s \parallel N_{n+1})$ as the next authenticating gene, then it makes sure that X , Y , Z, N have not been modified during the course of transportation by checking *Z*, if everything was right, the user would be recognized as a legitimate one, or his request would be denied.

Note $H(\bullet)$ here is a secure hash function, and $H^2(\bullet)$ is a duplicate hash function. Readers should be referenced to [8], if he wants to know more about this protocol.

2.2 Security Analysis with Respect to SDPA

Although SDPA is an improved scheme from SAS, and his security has got a great enhancement. We can still find a hidden trouble through the following simulation:

1) 2) The same as the phrase of user's identity authenticating in SDPA;

3) The adversary intercepts N_n and $M_n \oplus H^2(s \parallel N_n)$ from the server, then he (or she) substitutes $M_n \oplus H^2(s \parallel N_n)$ for a random equal length string r, and then sends N_n and r to the user;

4) *U* firstly computes $M_n' = H^2(s \parallel N_n) \oplus r$, then he (or she) chooses N_{n+1} randomly and computes: $X = H(s \parallel N_n) \oplus H(M_n)$, $Y = H^2(s \parallel N_{n+1}) \oplus H^2(M_n)$, $Z = H(H^2(s \parallel N_{n+1}) \parallel N_{n+1})$, $N = N_{n+1}$, and sends all of them to the server;

5) The adversary intercepts X, Y, Z, N , we can see from previous: $X = H(s \parallel N_n)$ $\oplus H(H^2(s \parallel N_n) \oplus r)$, $Y = H^2(s \parallel N_{n+1}) \oplus H^2(H^2(s \parallel N_n) \oplus r)$, also he (or she) knows N_n, N_{n+1} and r, so it's possible for him to get s with dictionary attack.

It's easy to get that this drawback lies to that N_n and N_{n+1} are transported in plaintext, if they were protected, this attack would disappear. The following is our own augmented scheme.

2.3 Augmented SDPA (ASDPA)

Register Phrase

1) A user *U* chooses one password *s* and a random number N_0, then computes $H(s)$ and $H^2(s \| N_0)$, and sends his (or her) ID, N_0, $H(s)$ and $H^2(s \| N_0)$ to identity authenticating server *S* via a secure channel;
2) The server *S* gets the user's ID, N_0, $H(s)$ and $H^2(s \| N_0)$ saved, note that $H^2(s \| N_0)$ here is used as *U*'s authenticating initiator.

Identity Authenticating Phrase

1) User *U* sends his request to the server;
2) Server S randomly chooses M_n and computes $M_n \oplus H^2(s \| N_n)$, then sends $N_n \oplus H(s)$ and $M_n \oplus H^2(s \| N_n)$ to the user;
3) *U* gets N_n and $H^2(s \| N_n)$ from $N_n \oplus H(s)$ and *s*, and it is easy for him to get M_n, then he randomly chooses N_{n+1} and computes: $X = H(s \| N_n) \oplus H(M_n)$, $Y = H^2(s \| N_{n+1}) \oplus H^2(M_n)$, $Z = H(H^2(s \| N_{n+1}) \| N_{n+1})$, $N = N_{n+1} \oplus N_n$, and sends them to the server together;
4) The server S calculates $H(X \oplus H(M_n))$ to authenticate the user, computes $Y \oplus H^2(M_n)$ to get $H^2(s \| N_{n+1})$ as the next authenticating gene, computes $N \oplus N_n$ to get N_{n+1}, then it makes sure that *X*, *Y*, *Z*, *N* have not been modified by checking *Z*, if everything was right, the user would be recognized as a legitimate one, or his request would be denied.

Compared to SDPA, ASDPA needs the server to store one more item $H(s)$ to protect N_n and N_{n+1} from being eavesdropped during the course of transportation, moreover, it needs the user to do one more hash computation and two more Xor operation, fortunately, the protocol turned to be more secure, so it is deserved. Of course, it is suitable to be used in our electronic payment protocol to authenticate user's identity.

3 Identity Based Cryptosystem

In 1984, Shamir [9] asked for a public key encryption scheme in which the public key can be arbitrary string, his original motivation is to simplify the certificate management in e-mail system, that is the sender can encrypt the mail with the receiver's e-mail address, and only the corresponding receiver can decrypt this mail and read it. Unfortunately, this idea hardly got any advancement until 2001, when D. Boneh constructed an Identity Based cryptosystem with Weil-Pairing in [10]. Compared to PKI, Identity-Based encryption (IBE) has a lot of advantages in key using and management. Firstly, the user's public key is his own identity (ID), such as his name, e-mail address and so on. As long as you know somebody's ID, you can sent message to him securely and verify his signature conveniently. Moreover, there is no need to store the ID in your computer, since it's so easy to be remembered. On the other hand, for mobile m-commerce's disadvantage, it's impossible to store user's certificate in mobile device for a long time as well as download it in real time, so IBE is the most perfect solution to encrypt and sign in mobile device.

While so far, most of the identity based signature schemes are implemented with Weil-Paring or Tate-Pairing, whose computational load is a little more heavy [11, 12], so it's not fit for mobile device, forasmuch, a new identity based signature scheme under elliptic curve cryptosystem (ECC) is put forward, noted as IBECDSA (Identity Based Elliptic Curves Digital Signature Algorithm).

Setup. KGC chooses a secure elliptic curve E with a base point P, whose order is a prime number n, and a random a satisfied $a < n$, which it keeps secretly. Define one secure hash function $H:\ \{0,1\}^* \to \mathbb{Z}_n{}^*$. Then publishes the common parameters params={ $E, n, P, aP, H(\bullet)$ }.

Extract. A user submits his (or her) ID information and authenticate himself (or herself) to KGC, then KGC randomly chooses $r \in_R \mathbb{Z}_n$ and computes $s^* = a \times H(ID)$, $R = rP = (x_r, y_r)$ and $x_0 = x_r \bmod n$, finally it sends $\{s = s^* x_0 + r, R\}$ to the user as private key.

Signing Protocol. Suppose the signer's private key is $\{s, R\}$. He (or she) randomly chooses $k \in_R \mathbb{Z}_n$ and computes $z = kP, e = H(m \| z)$, $S = se + k$, then $\{m, e, S, R\}$ is the signature of the message of m.

Verification. Suppose ID_s is the signer's ID and $R = (x_r, y_r)$. The verifier computes $x_0 = x_r \bmod n$, $Q = H(ID_s) \times aP$ and $z' = SP - e(x_0 Q + R)$, and he accepts the signature if the following equation holds: $e = H(m \| z')$.

Theorem. If Schnorr signature [13] is unforgeable, so is IBECDSA.

Proof: Now suppose $\{m, e, S, R = (x_r, y_r)\}$ is a valid forged IBECDSA signature, and its verifying key is ID_s. It is obvious that $\{m, e, S\}$ is a valid schnorr signature with verifying key $x_0 \times H(ID_x) \times aP + R$, where $x_0 = x_r \bmod n$. That is to say, if the adversary can forge a valid IBECDSA signature, he can also forge a valid

Schnorr signature, describing it in another way, if Schnorr signature is unforgeable, so is IBECDSA.

Since ECC can get the same security as other public key cryptosystem with a shorter key, it can just be used to make up the disadvantage of mobile device's limited storage and computational resource, so this scheme is suitable for wireless network.

4 Electronic Payment Scheme

Recently, the study on electronic coins is mainly focused on off-line electronic payment. That is to say, there is no need for bank or a third party's help in the course of paying, just as our real cash system. Of course, in order to implement a relatively perfect off-line electronic payment system, it's necessary for the electronic coin to be divisible. We also adopt a binary approach as [2, 3, 14] did to achieve it.

4.1 Divisible Electronic Payment Scheme

Just as what we said before, almost all of the mobile devices have some fixed characters (FC) which can be distinguished easily and also the mobile device which is used in electronic business hardly changes, so the user can have the FC of his mobile device as his public key, therefore the verifier can believe that the mobile device belongs to the user as long as the signature shown by the user is verified correctly. Since that, we can have electronic coin, mobile device and user tied together.

There are four main participants in our payment system: Arbitrage, Bank, Consumer and Shop. The arbitrage is a trusted third party, and it is responsible for authenticating the FC of consumer's mobile device and generating private keys for the consumers corresponding to the FC. Here we can see that the mobile device and the consumer are tied together with the help of Arbitrage. If the consumer wants to withdraw electronic coin from bank, the bank will return him one including the information of his mobile device's FC, so the coin is associated to the mobile device. In this way, only the user whom the mobile device belongs to can use the coin legitimately. During the course of payment, the consumer needs to show a valid signature to the shop, after verifying it, the shop believes that the cash belongs to the user and agrees to accept it. The details of the scheme is described as follows:

Register. The consumer submits his (or her) ID information and mobile device's FC to the Arbitrage, after authenticating and storing them, the Arbitrage responds with $Sig_A(H(FC))$ and $\{s, R\}$, the latter is used as the consumer's private key.

Open an Account. The consumer submits his mobile device's FC and the Arbitrage's signature $Sig_A(H(FC))$ to the bank, after verifying the signature, the consumer is asked to process ASDPA register, then he will get a legitimate account from the bank.

Withdraw. The consumer demonstrates to the bank that he is the right person whom the account belongs to with ASDPA. If it is true, the bank sends a signature

$Cash = Sig_B(Value \| FC \| timestamp)$ to the consumer, and then the consumer will verify the signature to ensure the coin's validity. Note here the coin is worth $Value$.

Payment. The consumer gets a challenge c from the shop and responds $\{FC, C = Sig_{FCC}(c-1)\}$ to it, the shop will believe that the mobile device with fixed character FC belongs to the consumer by verifying the signature, because only the right person can get the private key corresponding to the FC. Then the shop will verify $Cash = Sig_B(Value \| FC \| timestamp)$ to determine whether to deal or not. Finally, the consumer will send $Coin = Sig_{FCC}(n_{0...i} \| ... \| n_{0...j} \| Cash \| Date)$ to the shop, where $n_{0...i} + ... + n_{0...j}$ is the total expenditure this time. After verifying *Coin*, the shop accepts it.

Deposit. The shop sends the consumer's FC, *Cash* and *Coin* to the bank, if the cash is valid, the bank will find the record of this cash's history consumption, it should make sure that the shop is not double-depositing and has not violated the two rules. If every thing is right, the bank accepts this deposit and keeps *Coin* = $Coin = Sig_{FCC}(n_{0...i} \| ... \| n_{0...j} \| Cash \| timestamp)$ in record, or the shop's FC and history consumption records will be sent to the Arbitrage, who will disclose his identity.

4.2 Security and Efficiency Analysis

As an electronic payment scheme, it's necessary to have the following four properties: Unforgeability, Untracebility, no double-spending, no double-deposit. Now, we'd like to show that our system satisfies it.

Unforgeability. Electronic coin is a signature of the bank, which we have proven before is unforgeable, so the coin is also unforgeable. In addition, your valid electronic coin may be hold by other person, while he has not the private key corresponding to you mobile device's FC, so he cannot use it legitimately.

Untracebility. In our payment system, including opening an account, withdraw, paying and deposit, we can see nothing about user's information except for his mobile device's FC, while on condition that the arbitrage is trusted, it's impossible for anybody to associate the mobile device's FC of one user to his identity, so our payment system has the property of untraceability.

No Double-Spending. When the shop goes to deposit, the bank will make sure that the shop has not violated the two rules, because it has all of the used nodes corresponding to the cash recorded. Or, the shop's FC and history consumption record will be sent to the Arbitrage, who will disclose this shop's identity. That is to say, our system can prevent the consumer from double spending.

No Double-Deposit. Since there is time information in consumer's *Coin*, if the shop double deposit the coin, there must be another one having been recorded by the bank, so by comparing the timestamp in the two coins, it can insure that it is shop's double-deposit, rather than consumer's double-spending.

In addition, there is information of *Cash* in the *Coin*, so it is impossible for the shop to deposit the *Coin* by associating it to another *Cash*.

In the following, we will show that this system is efficient and fit for M-commerce.

1) In our system, both the consumer and shop can complete the payment with only twice identity based signing and verifying operation, so the computational load is light.

2) The *Cash* and the *Coin* are both IBECDSA signature, which is based on ECC. While currently, ECC can get sufficient security level with a key which is only 160 bits, so the IBECDSA signature can save a lot of storage capacity and network bandwidth.

In one word, our system has several advantages, which just make up the limitation of wireless communication. So undoubtedly, it's suitable for M-commerce.

5 Conclusion

Since there are a few limitations in mobile communication, it's hard to transplant so many protocols, which run perfectly in fixed terminals, to wireless network, so after fully considering the characteristic of mobile device, an efficient and divisible electronic payment scheme is proposed here. In the scheme, we achieve electronic coin's divisibility with binary tree approach, while the consumer can give the tree's node to the shop directly, rather than compute any other information. Otherwise, only several times IBECDSA signature is used in our scheme, so it has a lot of advantages in computational load, storage needed and network load, which just make up the limitation of wireless network. Also we have improved the SDPA dynamic password identity authentication scheme, and the augmented scheme is against man-in-middle attack. Meanwhile, a portable identity based signature scheme is proposed in this paper.

References

1. Brands, S.: Untraceable Off-Line Cash in Wallets with Observers. In: Proc. Crypto '93. Lecture Notes in Computer Science, Vol. 773. Springer-Verlag, Berlin Heidelberg New York (1994), 302–318
2. Ferguson, N.: Single Term Off-line Coins. In: Proc. EUROCRYPT '93. Lecture Notes in Computer Science, Vol. 765. Springer-Verlag, Berlin Heidelberg New York (1994), 318–328
3. Tatsuaki Okamoto: An Efficient Divisible Electronic Cash Scheme. In: Proceedings of the 15th Annual International Cryptology Conference on Advances in Cryptology. Lecture Notes in Computer Science. Springer-Verlag, London (1995), 438–451
4. Huang, Z., Chen, K.F., Zheng, D.: Efficient Electronic Payment Fit for Wireless Networks. Chinese Journal of Electronics. Vol. 13(2), Apr. 2004
5. Manjula Sandiringama, Akihiro Shimizu, Matu-tarow Noda: Simple and Secure Password Authentication Protocol. IEICE Trans Comm., 2000, E83-B(6), 1363-1365

6. Ku, W. C. , Chen, C. M.: Cryptanalysis of A One Time Password Authentication Protocols. In: Proceedings of the 2001 National Computer Symposium. Taiwan, Dec. 2001, 17046–17050
7. Lin, C. L., Sun, H. M., Hwang, T.: Attacks And Solutions on Strong-password Authentication. In: IEICE Transactions on Communications. Vol. E84-B(9), Sept. 2001, 2622–2627
8. Yuan, D., FAN, Z.P.: A Secure Dynamic Password Authentication Scheme. Journal of Sichuan University. Natural science edition. Vol. 39(2), Apr. 2002
9. Shamir: Identity-based Cryptosystems And Signature Schemes. In: Proceedings of CRYPTO'84. Lecture Notes in Computer Science, Vol. 196. Springer-Verlag, Berlin Heidelberg New York (1985), 47–53
10. Boneh, D., Franklin, M.: Identity Based Encryption from The Weil Pairing. In: Crypto 2001. Lecture Notes in Computer Science, Vol. 2139. Springer-Verlag, Berlin Heidelberg New York, 229–231
11. Barreto, P., Kim, H.Y., Lynn, B., Scott, M.: Efficient Algorithms for Pairings Based Cryptosystems. Advances in Cryptology-Crypto 2002. Lecture Notes in Computer Science, Vol. 2442. Springer-Verlag, Berlin Heidelberg New York (2002), 354–368
12. 12. Galbraith, S. D., Harrison, K., Soldera, D.: Implementing the Tate Pairing. In: Algorithmic Number Theory Symposium-ANTS-V. Lecture Notes in Computer Science, Vol. 2369. Springer-Verlag, Berlin Heidelberg New York (2002), 324–337
13. Schnorr, C. P. : Efficient Signature Generation for Smart Cards. Journal of Cryptology. Vol. 4(3), 1991, 161–174
14. Tatsuaki Okamoto, Kazuo Ohta: Universal electronic cash. In: CRYPTO '91. Lecture Notes in Computer Science, Vol. 576. Springer-Verlag, Berlin Heidelberg New York (1992).

Speech Authentication by Semi-fragile Watermarking

Bin Yan[1], Zhe-Ming Lu[1], Sheng-He Sun[1], and Jeng-Shyang Pan[2]

[1] Department of Automatic Test and Control, Harbin Institute of Technology
P. O. Box 339, 150001 Harbin, P.R. China
yanbinhit@hotmail.com, zhemingl@yahoo.com
[2] Department of Electronic Engineering
National Kaohsiung University of Applied Sciences
415 Chien-Kung Road, Kaohsiung 807, Taiwan
jspan@cc.kuas.edu.tw

Abstract. This paper proposes a semi-fragile speech watermarking scheme by quantization of Linear Prediction (LP) parameters, i.e. the inverse sine parameters. The watermark decoding performance is analyzed by modelling the parameters estimation error as Laplace distributed noises. The watermark detection threshold is derived according to the requirement of error probability and expected Signal to Noise Ratio (SNR). Experiments show that the proposed watermarking scheme is robust against amplitude scaling and semi-fragile to white noise addition, and thus suitable for speech authentication.

1 Introduction

In network environment, the multimedia contents transmitted through Internet may be intercepted and modified by adversaries; intruders may modify the multimedia database connected to Internet. If some decisions are to be made based on these multimedia contents, the decision maker must ensure that the information is authenticated. In military commanding systems, every received speech command should be authenticated. Fragile watermark has provided us such mechanisms [1]: a secure mark is embedded into the host media, modifications or substitution of the multimedia content will be identified by checking the existence of the mark after reception. Fragile watermark can also be combined with a robust watermark to form a multipurpose watermark [2]; the fragile watermark indicates whether there are modifications while the robust watermark provides the copyright information. Current speech watermarking techniques can be classified as:

– **Altering phase information** [3, 4]: These schemes uses the fact that the human auditory system is less sensitive to the changes in phase than in amplitude, all-pass filters are used to alter the phase information.
– **Spread spectrum** [5]: The watermark information is modulated using Direct Sequence Spread Spectrum/Binary Phase Shift Keying and then embedded into the residual of the speech signal after inverse filtering.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 497–504, 2005.

- **Parametric modelling** [6, 7]: The human articulatory system can be modelled as AR model driven by the stimulating signal. Gurijala et al. [6] proposes modifying the AR model parameters indirectly to embed the robust watermark, non-blind detection is required. Hatada et al. [7] proposes embedding watermark using vector quantization of LSP (Line Spectrum Pair) parameters.

In authentication applications, a fragile watermark should be embedded. Wu et al. [8] presented a fragile watermarking scheme by quantization of DFT coefficients in log scale. In [9], Lu et al. combined watermarking with the CELP (Code Excited Linear Prediction) speech coding process for authentication of compressed speech by CELP-typed coders, this authentication scheme is applicable only to compressed speech. The work presented here is the extension work in [9], and it aims at providing authentication of speech signal that is robust against amplitude scaling. The fragility of the watermark can be controlled by specifying the detection threshold according to the expected SNR.

2 Basic Embedding and Decoding Scheme

Watermarks are embedded into speech signal by modification of its parametrical representation, this work is different from Gurijala's parametrical scheme in that here the inverse sine parameters derived from LP coefficients are directly modified, so that the stability of the AR model can be guaranteed. In addition, for the purpose of authentication, the original speech signal is not required. Basic structure of the proposed watermarking system is shown in Fig. 1. Let $c[n]$ be the speech sample with index n, $\boldsymbol{R}_{\mathrm{C}}$ be the short-term autocorrelation matrix and $\boldsymbol{r}_{\mathrm{C}} = (r_{\mathrm{C}}[1], r_{\mathrm{C}}[2], \cdots, r_{\mathrm{C}}[P])^T$, where $r_{\mathrm{C}}[\eta] = \mathcal{L}\{c[j]c[j-\eta]\}$ and $\mathcal{L}(\cdot)$ denotes the time average of one realization of the WSS ergodic stochastic process. The host speech signal is first segmented into non-overlapping frames, here we consider only one frame of speech with L_{f} samples, the estimated LP coefficients are $\boldsymbol{a} = -\boldsymbol{R}_{\mathrm{C}}^{-1} \cdot \boldsymbol{r}_{\mathrm{C}}$, where $\boldsymbol{a} = \{a_i\}_{i=1}^{P}$. The residual signal of this frame is

$$e[n] = c[n] + \sum_{k=1}^{P} a[k]c[n-k] \tag{1}$$

where P is the order of AR model. These estimated LP coefficients cannot be quantized directly because the pole locations of AR model can't be controlled by LP coefficients, so the stability of watermarked AR model can't be guaranteed. In speech coding, the LP coefficients are usually transformed into Log Area Ratio (LAR) or Inverse Sine (IS) parameters to guarantee the stability and to reduce the spectral sensitivity. The LP coefficients are first transformed into Reflection Coefficients (RC) $\{\kappa_i\}_{i=1}^{P}$ and then the IS parameters $\{g_i\}_{i=1}^{P}$ by $g_i = \frac{2}{\pi}\sin^{-1}(\kappa_i), i \in \{1, 2, \cdots, P\}$. The modification to the IS parameters are based on odd-even modulation [1]

$$\hat{g}_i = \left\lfloor \frac{g_i + (1-w_i)\Delta_i)}{2\Delta_i} \right\rfloor \times 2\Delta_i + w_i\Delta_i, \quad i \in \{1, 2, \cdots, P\} \tag{2}$$

where $\lfloor\cdot\rfloor$ denotes the floor function, $\mathit{\Delta}_i$ is the quantization step for the i-th IS parameter, $w_i \in \{0,1\}$ is the watermark bit to be embedded. The modified IS parameters $\{\hat{g}_i\}_{i=1}^{P}$ are inverse transformed into LP coefficients $\hat{\boldsymbol{a}} = \{\hat{a}_i\}_{i=1}^{P}$. The residual from the LP analysis stage (1) is used to synthesize the watermarked speech signal $\hat{c}[n] = -\sum_{k=1}^{P}\hat{a}[k]\hat{c}[n-k] + e[n]$. In the watermark decoding process, the watermarked and possibly attacked signal $\tilde{\boldsymbol{c}} = \{\tilde{c}_i\}_{i=1}^{L_{\mathrm{f}}}$ is analyzed to get $\tilde{\boldsymbol{a}} = -\tilde{\boldsymbol{R}}_{\mathrm{C}}^{-1}\cdot\tilde{\boldsymbol{r}}_{\mathrm{C}}$. Decoding of watermark bits are performed according to

$$\tilde{w}_i = \left\lfloor \frac{\tilde{g}_i}{\mathit{\Delta}_i} + \frac{1}{2} \right\rfloor \pmod 2, \quad i \in \{1,2,\cdots,P\} \tag{3}$$

where $\{\tilde{g}_i\}_{i=1}^{P}$ are IS parameters from $\tilde{\boldsymbol{a}}$. Tamper Assessment Function (TAF) is calculated according to $\mathrm{TAF}(\boldsymbol{w},\tilde{\boldsymbol{w}}) = \frac{1}{N_{\mathrm{f}}\times P}\sum_{i=1}^{P}\sum_{j=1}^{N_{\mathrm{f}}} w_{ij}\oplus\tilde{w}_{ij}$ [1], where N_{f} is the number of speech frames involved in authentication. The detection statistic is chosen as $D = N_{\mathrm{f}}\times P - \sum_{i=1}^{P}\sum_{j=1}^{N_{\mathrm{f}}} w_{ij}\oplus\tilde{w}_{ij}$. Based on the Probability Density Function (PDF) of $\{\tilde{g}_i\}_{i=1}^{P}$, an appropriate threshold $\mathcal{T}$ is selected, if $D > \mathcal{T}$, then we can conclude that the speech signal is not modified.

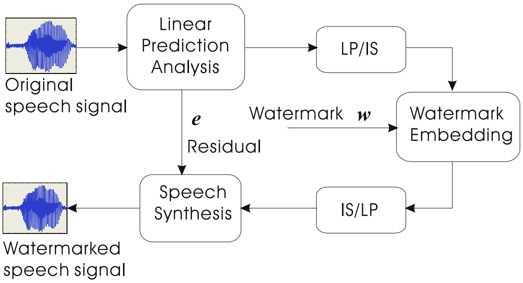


Fig. 1. Watermark Embedding

3 Decoding Performance Analysis

Due to their statistical nature, the estimated LP coefficients $\{\tilde{a}_i\}_{i=1}^{P}$ are different from the watermarked LP coefficients $\{\hat{a}_i\}_{i=1}^{P}$ even though there were no attacks. The estimation error of IS parameters can be modelled as additive noise on $\hat{g}$, it is found through the experiment and distribution fitting that Laplace distribution is appropriate to model the estimation error. Let $g_{\mathrm{E}} = \tilde{g} - \hat{g}$ be the estimation error of IS parameters, the histogram and empirical PDF of g_{E} are shown in Fig. 2. Let $\mathcal{U}_0$, $\mathcal{U}_1$ be the codebooks to embed watermark bit 0 and 1 respectively. Without loss of generality, we assume $\hat{g} = 0$ and $\hat{g} \in \mathcal{U}_0$, the noise variance σ^2 is also assumed to be independent of the quantization step $\mathit{\Delta}$. The PDF of $\tilde{g}$ is $p(\tilde{g}) = \frac{1}{\sqrt{2\sigma^2}}\exp\left(-\sqrt{\frac{2}{\sigma^2}}|\tilde{g}|\right)$. The cumulative distribution function and its complement are $F\left(\tilde{g};\sigma^2\right) = \frac{1}{2}\left(1 + \mathrm{sgn}(\tilde{g})\left[1 - \exp\left(-\sqrt{2/\sigma^2}\,|\tilde{g}|\right)\right]\right)$

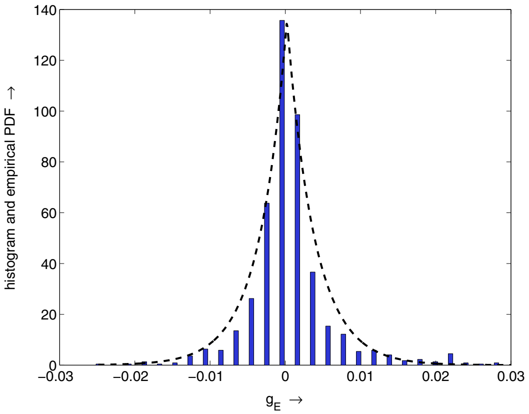


Fig. 2. Histogram of g_{E} and Laplace PDF with parameters estimated from g_{E} using Maximum Likelihood estimation

and $Q_L\left(\tilde{g};\sigma^2\right) = 1 - F\left(\tilde{g};\sigma^2\right)$. When watermark bit $w = 0$ is embedded, the probability distribution of $\tilde{g}$ is

$$p(\tilde{g}|w=0) = \sum_k p\left(\tilde{g}|u_{0,k}\right) \cdot P\left(u_{0,k}\right) \tag{4}$$

$$= \sum_k \frac{1}{\sqrt{2\sigma^2}} \exp\left(-\sqrt{\frac{2}{\sigma^2}}|\tilde{g} - u_{0,k}|\right) \cdot P\left(u_{0,k}\right) \tag{5}$$

where $u_{0,k}$ is the k-th codeword in codebook $\mathcal{U}_0$. Similar expression under condition $w = 1$ can also be obtained. The decoding process is based on Maximum-Likelihood decoding [10]: $\hat{w} = \arg\max_{w\in\{0,1\}} p(\tilde{g}|w)$. Here we use the fact that $p(\tilde{g}) \approx 0$ for $|\tilde{g}| > 2\Delta$, such that only adjacent cells are involved in this comparison, under this assumption the decoding error probability is:

$$P_{\mathrm{e}} = \Pr\left\{\tilde{g} \in \mathcal{R}_1 | \hat{g} \in \mathcal{R}_0\right\} \approx 2 \times \left[Q_L\left(\frac{\Delta}{2};\sigma^2\right) - Q_L\left(\frac{3\Delta}{2};\sigma^2\right)\right] \tag{6}$$

$$= \exp\left(-\frac{1}{\sqrt{2}}\frac{\Delta}{\sigma}\right) - \exp\left(-\frac{3}{\sqrt{2}}\frac{\Delta}{\sigma}\right) \quad , \tag{7}$$

where $\mathcal{R}_0$ and $\mathcal{R}_1$ are cells associated with codebook $\mathcal{U}_0$ and $\mathcal{U}_1$ respectively. For $\Delta = 0.0156$ and $\sigma \approx 0.0047$, P_{e} is calculated to be 9.5% according to (6), the experimental results is on average TAF = 8.1%, the small difference is due to the inaccuracy of the probability model and the randomness of the watermark sequence used in each experiment.

4 The Effects of Attacks

Amplitude Scaling. It is desired that speech authentication algorithms should be robust against amplitude scaling, unfortunately almost all quantization watermarking based authentication schemes have suffered from this, the reason is that amplitude scaling will cause the mismatch of quantization steps between the watermark embedder and decoder. In the proposed scheme, the algorithm works on the model of vocal tract, the power of the speech signal is left to be controlled by the power of the glottal excitation. In the vocal tract model $H(z) = 1/\left(1 + \sum_{i=1}^{P} a_i z^{-i}\right)$, we don't need to consider the power term. When the amplitude of the speech signal is amplified by a factor α, the estimated a_i remains unchanged, while the amplitude of the residual signal will be amplified by the factor α. An experiment is designed to test the above stated property, scaling factors from 0.1 to 2 are used to scale the amplitude of the watermarked speech signal, the extracted IS parameters are identical to the case without amplitude scaling.

Addition of White Gaussian Noise. Figure 3 shows the effects of WGN attack on TAF values. The underlying mechanism is that WGN affects the estimation accuracy of LP parameters and hence the IS parameters, especially in unvoiced regions. In current scheme, we don't consider the difference between voiced and unvoiced speech. Actually, the long tails in the empirical distribution of g_{E} was due largely to the estimation error in the unvoiced region. If the application requires the algorithm to tolerate more WGN attacks, then the watermarking algorithm should be modified to consider embedding watermark only in the voiced region. The experimental results in Fig. 3 will be used to determine the watermark detection threshold. In applications where the expected WGN power is large, the detection algorithm should lower its detection threshold to lower the detection error.

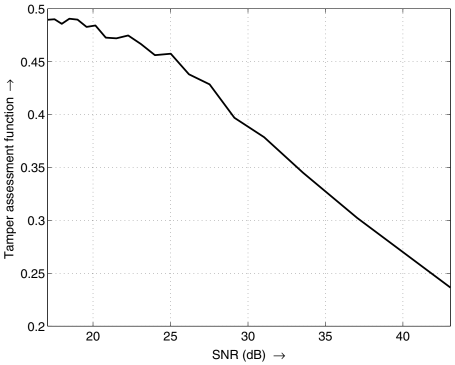


Fig. 3. Tamper assessment function for white Gaussian noise attack

5 Watermark Detection and Content Authentication

The extracted watermark bit sequence is fed into the watermark detector. The detection threshold is derived according to the application requirements where the detection error probability and expected SNR are specified. The TAF value gives us the estimate of $\Pr\{\hat{w} \neq w\}$, where $\hat{w}$ is the extracted watermark bit. In the presence of WGN attacks, the relation between TAF and SNR as shown in Fig. 3 can be well approximated by a straight line where the SNR ranges from 25 dB to 45 dB. Let this function be $\text{TAF} = g(\text{SNR})$. Then we have $\Pr\{\hat{w} = w; \text{with watermark}\} \approx 1 - g(\text{SNR})$. When parts of the speech signals are replaced, we have $\Pr\{\hat{w} = w; \text{no watermark}\} = 0.5$. Let the hypothesis be $\mathcal{H}_0$: No watermark; and $\mathcal{H}_1$: Watermarked with $\boldsymbol{w}$. Then the probability that k corresponding bits are identical between original and extracted watermark sequences is $P(k|\mathcal{H}_0) = \mathcal{B}\left(N_{\text{w}}, 0.5\right)$, and $P(k|\mathcal{H}_1) = \mathcal{B}\left(N_{\text{w}}, 1 - g(\text{SNR})\right)$, where $\mathcal{B}(\cdot)$ denotes binomial distribution, N_{w} is the length of the watermark sequence. For sufficient large N_{w}, binomial distribution can be approximated by Gaussian distribution, so we have $P(k|\mathcal{H}_0) = \mathcal{N}\left(N_{\text{w}}/2, N_{\text{w}}/4\right)$, and $P(k|\mathcal{H}_1) = \mathcal{N}\left(N_{\text{w}}\left(1 - g(\text{SNR})\right), N_{\text{w}}\left(1 - g(\text{SNR})\right) \cdot g(\text{SNR})\right)$, where $\mathcal{N}(\mu, \sigma^2)$ denotes Gaussian distribution with mean μ and variance σ^2. Then the false alarm and missed detection probabilities are

$$\begin{aligned} P_{\text{FA}} &= \int_{\mathcal{T}}^{\infty} P(k|\mathcal{H}_0) \cdot dk = Q\left(\frac{\mathcal{T} - N_{\text{w}}/2}{\sqrt{N_{\text{w}}}/2}\right) = \alpha \\ P_{\text{MD}} &= \int_{-\infty}^{\mathcal{T}} P(k|\mathcal{H}_1) \cdot dk = 1 - Q\left(\frac{\mathcal{T} - N_{\text{w}}(1 - g(\text{SNR}))}{\sqrt{N_{\text{w}}(1 - g(\text{SNR})) \cdot g(\text{SNR})}}\right) = \beta \end{aligned} \tag{8}$$

where $Q(x) = \int_x^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$ is Q-function, k is considered as continuous variable. The detection threshold and authentication length are

$$\begin{aligned} N_{\text{w}}^{*} &= \left[\frac{Q^{-1}(1-\beta)\sqrt{g(\text{SNR}) \cdot (1 - g(\text{SNR}))} - 0.5 Q^{-1}(\alpha)}{g(\text{SNR}) - 0.5}\right]^2 \\ \mathcal{T}^{*} &= \frac{N_{\text{w}}^{*}}{2} + \frac{\sqrt{N_{\text{w}}^{*}}}{2} Q^{-1}(\alpha) \ . \end{aligned} \tag{9}$$

In the cases when any modification to speech should be identified, the term $g(\text{SNR})$ should be replaced with P_{e} in (6). Note that due to the non-stationary nature of speech signals, the relation $P(w \neq \hat{w}) \approx g(\text{SNR})$ is estimated from a long speech segment, so (9) is valid only for very small α and β, which is the case in practical applications. The security of the proposed watermarking system relies on the watermarking key K, which consist of (1) the watermark sequence $\boldsymbol{w}$, (2) the IS parameters selection key K_{s}, (3) the order of LP analysis P, and (4) the quantization steps for each IS parameter. Without knowledge of the watermarking key, the attackers are difficult to watermark the replaced speech segments. Figure 4 shows the histogram of detection statistic, which are well separated. In this experiment, 50 frames of a speech signal are involved in speech authentication, the detection threshold is $\mathcal{T} = 435$ according to (9). Several frames of the original speech signal and watermarked speech signal are plotted and compared in Fig. 5. The informal subjective listening test indicates no audible distortion between the original and watermarked speech signals.

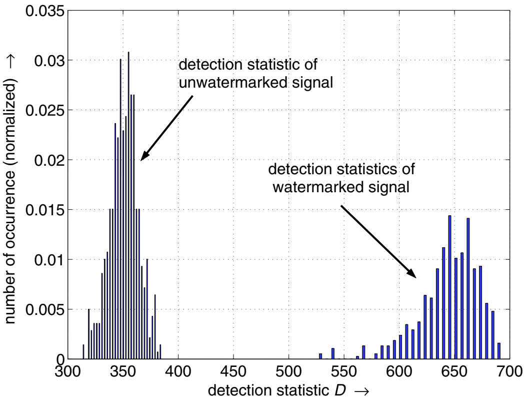


Fig. 4. Histogram of detection statistics

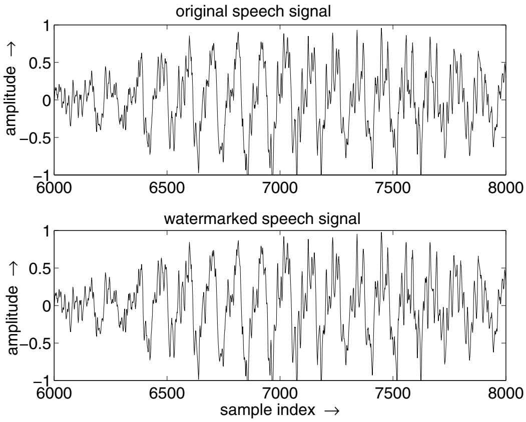


Fig. 5. Speech waveform without and with watermark

6 Conclusion and Future Work

This paper presented a semi-fragile watermarking technique for speech authentication. The proposed scheme is robust against constant amplitude scaling. The estimation error of IS parameters causes the watermark decoding error in the absence of attack, which can be compensated by statistical detection. The fragility of the watermark can be controlled by the watermark detection threshold, which is derived from probability of error requirement and expected SNR.

The quantization step of IS parameters Δ is selected by trail and error and subjective listening test. The algorithm on determining Δ using systematic approach is under test, A modified bit allocation algorithm can be used to achieve the 'transparency' requirement. Further research will also be focused on how to reduce the IS parameter estimation error, an Analysis by Synthesis (AbS) framework is adopted as a preliminary solution.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under grant 60272074 and the Spaceflight Innovation Foundation of China under grant [2002]210-6.

References

1. Kundur, D.: Multiresolution Digital Watermarking: Algorithms and Implications for Multimedia Signals. Ph. D. Thesis. Graduate Department of Electrical and Computer Engineering, University of Toronto. (1999)
2. Lu, Z. M., Xu, D. G. and Sun, S. H.: Multipurpose Image Watermarking Algorithm Based on Multistage Vector Quantization. IEEE Transactions on Image Processing. **14** (6) (2005) .Accepted for publication
3. Yardyimci, Y., Cetin, A. E., and Ansari, R.: Data Hiding in Speech Using Phase Coding. Eurospeech 97 **3** (1997) 1679-1682
4. Ciloglu, T. and Karaaslan, S. Utku: An Improved All-Pass Watermarking Scheme for Speech and Audio. International Conference on Multimedia and Expo. July 30-Aug. 2 (2000) **2** 1017-1020
5. Cheng, Q. and Sorensen, J.: Spread Spectrum Signaling For Speech Watermarking. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. May 7-11 (2001). 1337-1340
6. Gurijala, A. R., Deller, JR., J. R.: Speech Watermarking with Objective Fidelity and Robustness Criteria. Proceeding of Asilomar Conference on Signals, Systems, and Computers, Pacific Grove CA, Nov. (2003)
7. Hatada, M.,Sakai, T.,Komatsu, N. and Yamazaki, Y.: Digital Watermarking Based on Process of Speech Production. Proceedings of SPIE. **4861** (2002) 258-267
8. Wu, C. P. and Jay Kuo, C.-C.: Fragile Speech Watermarking Based on Exponential Scale Quantization for Tamper Detection. the IEEE International Conference on Acoustic, Speech and Signal Processing. (2002) 3305-3308
9. Lu, Z. M., Yan, B. and Sun, S. H.: Watermarking Combined with CELP Speech Coding for Authentication. IEICE Transactions on Information and systems. **E88-D** No.2. (2005) 330-334
10. Barni, M., Bartolini, F.: Watermarking Systems Engineering: Enabling Digital Assets Security and Other Applications. Marcel Dekker, Inc. New York. (2004)

On Engineering Smart Systems

E.V. Krishnamurthy[1] and V. Kris Murthy[2]

[1] Computer Sciences Laboratory
Australian National University, Canberra, ACT 0200, Australia
abk@discus.anu.edu.au
[2] School of Business Information Technology, R.M.I.T University
Melbourne, 3000, Australia
kris.murthy@rmit.edu.au

Abstract. A smart system exhibits the four important properties: (i) Interactive, collective, coordinated and efficient Operation (ii) Self -organization and emergence (iii) Power law scaling under emergence (iv) Adaptive. We describe the role of fractal and percolation models for understanding smart systems. A hierarchy based on metric entropy is suggested among the computational systems to differentiate ordinary system from the smart system. Engineering a general purpose smart system is not feasible, since emergence is a global behaviour (or a goal) that evolves from the local behaviour (goals) of components. This is due to the fact that the evolutionary rules for the global goal is non-computable, as it cannot be expressed as a finite composition of computable function of local goals for any arbitrary problem domain.

1 Introduction

Smart systems have no formal definitions. Defining smart systems is not easy, although understanding the properties of such a system is necessary to understand the properties of autonomous intelligent systems occurring in Nature. No suitable formal model is available for smart systems. We, therefore define smart systems as those systems having the following important properties:

1. 1nteractive, Collective, Coordinated and Efficient Operation
A smart system consists of a large number of components and can interact with the environment (hence called open). Also they collectively and cooperatively perform actions, coordinating their actions when there is competition, to obtain maximal efficiency.

2. Self- organization and Emergence
The total dynamic behaviour arising due to cooperation between different parts of the system can lead to a coherent behaviour of the entire system that can change by a large amount when the values of a control parameter changes by a small amount (sensitivity). These changes have similarities to the phase transition encountered in physics, e.g., change of a substance from a liquid to a gas, in which new properties emerge abruptly. These new properties of the system are not predictable, in advance from the properties of the individual interactions. In particular, under emergence, the many degrees of freedom arising due to its component parts collapse into a fewer new ones with a smaller number of globally relevant parameters.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 505–512, 2005.

3. Power-Law Scaling Property
Power -law scaling relationship arises for the newly emergent properties of the smart system.

4. Adaptive, Fault-Resilient and Flexible
Smart systems are flexible -they can self-modify their past behaviour and adapt to environmental changes, available resources, as well as, tolerating failures or non-cooperation of some of their components.

In this paper we restrict our consideration to the following issues:
1.What kind of models are suitable for understanding smart systems?
2.Is the behaviour of smartness analogous to the critical phenomenon in physics or percolation? Can we obtain suitable parameters to describe this phenomenon?
3. Is there a hierarchy of degree of smartness among computational systems?
4. Can we engineer smart systems to solve problems?

2 Conrad's Principles

According to Conrad [6] a general purpose computing system cannot have all of the three following properties:
1. Structural programmability (Algorithmizability).
2. High computational efficiency.
3. High evolutionary adaptability and flexibility.

Properties 1 and 2 are mutually exclusive. That is, we cannot have a high computational efficiency in a programmable system. Properties 1 and 3 are mutually exclusive in the region where maximum computational efficiency exists. That is, whenever maximal computational efficiency is required, the system should be highly adaptive rather than structurally programmable (algorithmic). Adaptation implies self-modification and long-range correlations exist among the widely separated different parts leading to self organization and emergence.

For example, a diseased heart or lung alters its structure (fractal dimension) for survival. Such an organization optimizes the consumption of oxygen and transport of oxygen in blood or the pumping rate, Leibovitch [16]. Nature always prefers changing shape and structure, as it is cheaper than new materials. Conrad's principles, lead to the fact that self- organization is not possible in an algorithmic system.In fact, any external order imposed on such systems can prevent self-organization, Boettcher and Percus [2], Camazine et al. [4].

3 Fractal, Percolation and Brownian Dynamics Models

A smart system requires three information processing features:

1. Information storage to preserve the information about the local state for arbitrarily long time.
2.Information transmission to propagate information over arbitrarily long distances.
3.Interaction of stored information with the received information.

The above three features require that any physical system supporting computation must exhibit arbitrarily large correlation lengths in space and time. Hence, percolation theory and fractal dynamics can explain the following aspects:

1. Cross scale interactions that arise among micro and meso scale processes and percolate to macro scale beyond a critical value.
2.The relationship among basic functional units at different spatial /time resolutions in terms of geometrical and statistical properties.
3. Many intractable algorithms exhibit phase transitions like phenomena having a critical point obeying a power law in this region, Hubermann and Hogg [13].
4. The probabilistic interaction with environment results in a change in the fractal dimensionality in space and time leading to a phase transition like behaviour and self organized criticality.
5. A system undergoing phase transition shows a large change in behaviour resulting in bifurcation having fractal features. Such examples arise in motor and sensory phase transitions in living systems; Bunde and Havlin [3], Leibovitch [16].

3.1 Fractal Geometry and Percolation

A smart system has an associated data domain or space. This space is usually the three dimensional space; it is called the geometric dimension of the smart system. When the system is placed in an environment, it does not communicate through its three dimensional volume, but through its surface area. Since the amount of communication is proportional to the surface area, a simple way to increase communication rate is to choose a fractal object that have a larger surface area compared to compact objects, Bunde and Havlin [3], Grimmett [11], Leibovitch [16], West [24],Wolfram [25]. Therefore in biology and chemistry, where surface phenomena play a crucial role, nature has preferred the evolution of fractal objects. Such a choice maximizes the interaction with the environment and has the following advantages:

1.Modification of surface area provides for adaptation and the survival.
2.Absence of characteristic length and long-range correlation result in scaling laws.

In heterogeneous chemistry, the structure and geometry of the environment at which the reaction takes place plays an important role; it alone can dictate whether a reaction will take place at all. The geometric parameter is the fractal dimension. In fact environmental interaction can change geometrical features and conversely geometrical feature modify the interaction; see chapter on “Fractal Analysis” by Avnir et al. in Bunde and Havlin [3].

Percolation theory, Grimmett [11] is concerned with the study of the diffusion / penetration of certain materials from an environment into the material of system placed in that environment through the boundary or surface of the system. It has been found experimentally that such a penetration is determined by a single parameter p. We call a particular value p = p(c) the "threshold” for percolation, if the pathways are infinite when p > p(c) and the pathways are nonexistent or have limited access within the material of the system when p < p(c). The value p(c) is called the critical point. Also we say that there is a percolation above p(c). The region above p(c) is called the supercritical phase and the region below p(c) is called the subcritical phase. Also when p = p(c) the infinite open cluster or pathway is called an incipient infinite cluster.

Percolation theory is studied using a simplifying assumption of an internal connected structure of the material of the system, either as a lattice or as a connected graph. Also it is studied as a "bond percolation" in which we study the percolation through the edges of the lattice or a graph; or as a "site percolation" in which we study the percolation through the lattice points or vertices of the graph or as a "mixed percolation" involving edges and vertices.

For example, consider a square lattice, where each site is occupied randomly with probability p or empty with probability (1-p). At low concentration p the occupied states are either isolated or forms small clusters. Two occupied sites belong to the same cluster if a path of nearest neighbour occupied sites connects them. When p is increased, the average size of clusters increases. At a critical concentration p(c) a large cluster appears which connects the opposite edges of the lattice. This cluster is called the infinite cluster since it diverges when the size of the lattice is increased to infinity. When p increases further the density of the cluster increases, since more and more sites become part of the infinite cluster, and the average size of the finite clusters decreases.

The behaviour of percolation is described by the following parameters.

1.The number of infinite pathways (clusters): 0 below p(c); 1 above p(c).
2.Percolation Probability: T(p)= 0 below p(c); T(p) > 0 above p(c).
3.Mean Cluster size: C(p) < infinity below p(c); C(p) = infinity above p(c).
4.Tail of finite clusters decays exponentially below the threshold and decays slower than exponentially above the threshold.

Scaling theory predicts a power law relationship for the parameters such as: percolation probability T(p), truncated mean cluster size C(p), number of clusters per vertex N(p), and correlation length L(p), Grimmett [11], Dorogovtsev, and Mendes [7]. The above parameters have the following power law relationships with respect to (p-p(c)) = x; T(p)= x exp.b; C(p)=x exp -g; N(p)= x exp (-1-a); L(p)= x exp -n. The exponent involved here is called a Critical Exponent. It is widely accepted that: 2-a = g+2b.

Unfortunately, the behaviour of T(p) and C(p) are still not well-understood. While p(c) depends explicitly on the type of lattice (e.g. In site percolation p(c)= 0.593 for square lattice and 0.70 for hexagonal lattice), all the critical exponents depend upon the dimension of the lattice, but not on the type of the lattice structure. This is based upon the hypothesis that critical phenomena manifest themselves over large length scales and local lattice structure is irrelevant over such distances. This property is known as the Universality of the critical exponent. Universality can be explained as below: Consider the class of all percolation processes on d-dimensional lattices having a finite vertex degree. It is believed that the nature of the phase transitions thereof is similar; in particular, all processes in this class are believed to have equal critical exponents.

3.2 Random Fractals and Brownian Dynamical Model

Random fractals can have a complicated topological structure. For example, they can be highly multiply connected and the topological structure can change dramatically, when a continuously varying parameter increases through a critical value resulting in a simply connected path, Falconer[10]; Grimmett[11]. Thus unforeseen new paths can emerge between points of interest, e.g., nest to the food when communication takes place among agents in an environment; the use of Brownian dynamical agents, in this context, has been extensively studied recently by Ebeling and Schweitzer [8].

4 Creating a Smart System

By definition,a smart system exhibits the properties of both the computational and the dynamic systems. Hence we need to study how these systems are interrelated through parameters such as entropy, fractal dimension and Lyapunov exponents. A dynamical system evolves with time. Nonlinear systems can have attractors of four kinds Leibovitch [16], Falcioni et al. [9]. Ott et al. [19], Wolfram [25].

1. Fixed or equilibrium points
2. Periodic orbits
3. Quasi periodic attractors
4. Chaotic or strange attractors

Small parameter changes in a non-linear system can lead to an abrupt change in behaviour, e.g bifurcation. Lyapunov exponent serves as an important invariant to quantify the behaviour of an attractor. A system with negative Lyapunov exponents imply stability and a system with the positive Lyapunov exponents implies instability and the presence of positive feedback, Camazine et al.[4]. Most classical dynamical systems are associated with regular periodic motion and the associated differential equations are solvable; hence they are called Integrable systems.

If a system is nonlinear it turns out to be nonintegrable and exhibits various degrees of irregular dynamics, Falcioni et al.[9]:

(i)Ergodicity, (ii) Mixing, (iii) K-flow or Chaos (iv) Non- equilibrium systems

Each of the above properties imply all the preceding ones e.g., within a chaotic region the trajectories are ergodic on the attractor and wander around the desired periodic orbit. Classical motion is chaotic if the flow of the trajectories in a given region of phase space has positive Lyapunov exponents (or positive metric entropy) that measure the rate of exponential separation between two neighbouring points in the phase space.

Chaos indicates hypersensitivity on the initial conditions. Also the system becomes inseparable (metric transitivity) and the periodic points are dense. That is the whole dynamical system is simply not a sum of parts and it functions as a whole leading to what is known as " Emergence", and the algorithmic independence among the subsystems is lost, Chaitin [5].

Positive entropy system produces algorithmically complex sequences that are incompressible and unpredictable. In general, the evolutionary rules of systems with zero entropy are predictable (or Lyapunov exponent equal to zero) and such rules are not predictable for positive entropy machines, Falcioni et al. [9], Ott et al.[19]. Well-structured objects (e.g., Context free grammars, regular grammars and serial-parallel orders) provide for easy description through functional rules and hence have zero metric entropy. The systems with zero metric entropy are "Turing or algorithmically expressible" with a finite set of evolutionary rules.

When metric entropy is positive, in the long run the recording of information for evolution increases unbounded and the evolution cannot be followed deterministically, unless a disproportionately long or infinite time is devoted to this task beyond a critical time t = T(c). As we approach T(c), the recording is critically slowed down - much like in phase transition phenomena. At this time the motion is chaotic and the forward and backward evolution are not reversible resulting in a spontaneous breakdown of time -reversal symmetry. Thus a phase -transition like situation arises between the Turing expressible (tractable) and Turing non-expressible (intractable) systems leading to a critical point behaviour.

Based on Metric-Algorithmic entropy we can classify the two major classes of machines, ordinary (O) and dissipative (P) machines based on metric entropy thus:

1. Ordinary or Zero Metric Entropy Machines(O)
These are Completely structured, Deterministic, Exact behaviour (or Algorithmic) Machines. This class contains, the machines in Chomskian hierarchy:
(i) Finite State machines: obeys regular grammar or type 3 grammar;
(ii) Push down-stack machines: obeys context-free grammar or type 2 grammar;
(iii) Linear bounded automata: obeys context sensitive or type 1 grammars;
(iv) Turing Machines that halt: obeys an unrestricted or type 0 grammar and
(v) Exactly integrable Hamiltonian flow machines.

These machines are in principle, information loss-less and instruction obeying; their outputs contain all the required information, as dictated by the programs.

2. Positive Metric Entropy Machines (P)
(i) Ergodic (ii) Mixing (iii) Chaotic. (iv) Non-equilibrium

Thus to create a smart system, we need to combine the zero and positive entropy machines. This results in a multi-fractal and hyper-chaotic systems with many positive Lyapunov exponents and bifurcations. The injection of positive entropy through chaotic (deterministic randomness) or stochastic parameters (statistical randomness) also results in the following advantages:
(i) Provides ergodicity in the search orbits,
(ii) Provides solution discovery capabilities (as in genetic programming) since chaotic orbits are dense with positive Lyapunov exponents, and two initially close orbits can separate exponentially from each other, Prigogine [20,21], Zak et al. [26], Koza et al.[14], Holland [12].

4.1 Quantifying Smartness

Quantifying the complexity of “Smartness” requires the evaluation of three important parameters: Entropy, Lyapunov exponents, and Fractal dimension. Although the different studies show the existence of a phase -transition like phenomena and power law relationships do hold, as yet, we cannot confidently predict the critical probability and power law exponents without experimentation for each individual complex problem. Scaling and universality are widely accepted experimentally in many different areas (including Biology, Medicine, Physics and Social sciences and in random networks (including the World-Wide-Web), Serugendo [23] resulting in the small–world phenomena. Scaling relationship corresponds to a power law behaviour over a wide range of control parameter; the exponent involved in this power law is called the critical exponent. In fact, Dorogovtsev and Mendes [7] prove that power law scaling is essential for fault resilience. Also percolation / phase transition models reflect the cooperative, and competitive behaviour among the microscopic objects. The modelling uses a suitable geometric structure with a local computation that results in a global change. Further, the percolation systems have the property 4, namely, adaptive and flexible behaviour to tolerate failures, since it deals with the formation paths among distant neighbours, even if some of these neighbours are non-cooperative. This leads to an emergent behaviour through self-organized criticality. *Thus it appears that all the four properties are not necessarily independent;* Rose and Lauder [22].

The four required properties of smart systems are difficult to realise within the time reversible evolutionary systems. Hysteresis, long term memory and time arrow play a role in turning the systems smarter. In fact, it is due to the time- arrow (ageing) and memory, the smart system remembers and orders the temporal events as earlier and later, without the explicit sequential addressing mode used in programming. This leads to a kind of self-awareness of past, present and future or a psychological time arrow. From the above arguments we see that the smart system lies between order and disorder; see Langton et al. [15], Wolfram [25]; Zak et al.[26].

4.2 Engineering Emergent Behaviour

The central question in designing smart systems is how to program the components so that the system as a whole self organizes. This is the basic question addressed in the design of Amorphous computers, Abelson et al., [1] and Spray computers, Mamei and Zambonelli [17]. For engineering applications this would require an approach analogous to controlling and coping with chaos, Ott et al. [19]. Since emergence is a global behaviour (or a goal) that evolves from the local behaviours (goals) of components, the evolutionary rules for arriving at the global goal is non-computable, since it cannot be expressed as a finite composition of computable deterministic function of local goals for any arbitrary problem domain. Thus the system we have designed may not be compliant to our demand. The only way seems to be the “Darwinian Evolution” as advocated in “Genetic programming (GP)”, Koza et al[14], to create emergent programs or rules by topologically crossing-over the various subtrees in a set of possible programs. However, we think that the Genetic programming requires a paradigm-shift by introducing the features of the evolution of human mind as advocated by Mithen [18]. This would allow general –purpose intelligence to be supplemented by multiple specialized intelligence units, each devoted to a special domain of behaviour, and working independently and cooperatively. Also, this would enable us to design problem-specific techniques by dynamically balancing exploration versus exploitation of discovered knowledge.

5 Conclusion

We described some important properties a smart system need to possess. Smart systems exhibit the properties of both the computational and dynamical systems and can undergo phase transition from the former to the latter. Although, phase transition-like model is suitable for understanding a smart system, we cannot yet confidently compute the required quantifying parameters - such as metric entropy, Lyapunov exponents, Fractal dimension, critical probability and power law exponents without experimentation for each individual complex problem. This can be attributed to the break-down of the algorithmic structure and the emergence of noncomputability. This means for each problem in each domain, we need to understand the functions, their algebraic structure and composition and how to reach a global goal from a set of local goals. Here, the evolutionary psychology model of human mind may provide an answer.

References

1. Abelson, H et al.: Amorphous computing, Comm ACM, Vol.43, 5(2000)74-82.
2. Boettcher. S and Percus, A: Nature's way of Optimizing, Artificial Intelligence, 119(2000) 275-286.
3. Bunde, A and Havlin,S.: Fractals in Science, Springer Verlag, New York(1994)
4. Camazine, S et al: Self-Organization in Biological Systems, Princeton University Press, Princeton (2002)
5. Chaitin, G.: Two-philosophical applications of Algorithmic Information Theory, Lecture Notes in Computer Science, 2731(2003) 1-10
6. Conrad, M: Molecular Computing, Advances in Computers, 31, Academic Press, New York (1990), 235-325
7. Dorogovtsev, S.N and Mendes, J.F.F, Evolution of Networks, Oxford University Press, Oxford (2003)
8. Ebeling, W and Schweitzer, F.: Self-organization,Active Brownian dynamics and biological Applications, Nova Acta Leopoldina, 88, No.332(2003)169-188
9. Falcioni, M., et al, Kolmogorov's legacy about entropy, Chaos and Complexity, Lecture Notes in Physics, 636, Springer Verlag, New York (2003) 85-108
10. Falconer, K.: Fractal geometry, Wiley, New York (2003)
11. Grimmett, G.: Percolation, Springer, New York(1999)
12. Holland, J.H.:Emergence-chaos to Order, Addison Wesley, Reading, Mass (1998)
13. Hubermann,B.A. and Hogg, T: Phase transitions in Artificial Intelligence Systems, Artificial Intelligence, 33(1987) 155-171
14. Koza,R et al., Genetic programming III, Morgan Kaufmann, San Francisco (1999)
15. Langton,C.E. et al.: Life at the Edge of chaos, in Artificial Life II, Addison Wesley, Reading, Mass.(1992) 41-91
16. Liebovitch, L.S.: Fractals and Chaos,Oxford University Press, Oxford (1998)
17. Mamei,M., and Zambonelli, F.: Spray computers: Frontiers of self organization and Pervasive computing, www.irit.fr/SMAC/Publications.html (2004)
18. Mithen, S., The Prehistory of Mind, Thames and Hudson, London (1999)
19. Ott, E. et al.,Coping with Chaos, John Wiley, New York (1994)
20. Prigogine,I.: From being to becoming,W.H.Freeman and Co, San Francisco (1980).
21. Prigogine,I.:Laws of Nature, Physica A 263(1999) 528-539
22. Rose,M.R and Lauder,G.V. Adapatation, Academic Press, New York (1996)
23. Serugendo,G.D.M.: Engineering Emergent Behaviour: A Vision, Lecture Notes in Artificial Intelligence, 2927, Springer Verlag, New York (2003) 1-7
24. West, B.J.: Fractal Physiology and chaos in Medicine, World Scientific, Singapore (1990)
25. Wolfram, S.: A New kind of Science, Wolfram Media Inc., Champaign, Ill (2002)
26. Zak, M, et al.:From Instability to Intelligence, Springer Verlag, New York (1997)

Distributed Web Integration with Multiagent Data Mining

Ayahiko Niimi[1], Hitomi Noji[2], and Osamu Konishi[1]

[1] Department of Media Architecture, Future University-Hakodate
116–2 Kamedanakano-cho, Hakodate 041–8655, Japan
{niimi,okonishi}@fun.ac.jp
[2] Goodwill Engineering
6–10–1 Roppongi Minato-ku, Tokyo 106–6137, Japan

Abstract. We proposed a technique for using multiagent technology in data mining intended for two or more text databases. In this paper, we discuss data mining method based on text (text mining), but our proposed method is not a method of specializing in text mining. First, we proposed data mining technique using multiagent technology. The proposed technique is applied to document databases, and discuss its results. In this paper, proposed data mining using multiagent was applied to information integration system on Web, and the effectiveness was verified. In the proposed method, the part of the database access agent was changed to the Web access agent. Also, mining agent was changed to the information extraction agent from the HTML file.

1 Introduction

In KES2003 and KES2004, we proposed a technique for using multiagent technology in data mining intended for two or more text databases. [1, 2] We applied our proposed approach to data mining from the document database, and discuss its problems. To apply proposed approach, we constructed only a minimum mounting which runs only UNIX local machine with process communications as agent communication and file system as black board model. It was confirmed to be able to switch the database and the data mining algorithm that used the constructed data mining system. We discussed data mining method based on text (text mining), but our proposed method is not a method of specializing in text mining.

In this paper, proposed data mining using multiagent was applied to information integration system on Web, and the effectiveness was verified. In the proposed method, the part of the database access agent was changed to the Web access agent. Also, mining agent was changed to the information extraction agent from the HTML file. The information integration on the Web page can be thought just like the information integration from the database. Similarly, it can be thought that the information extraction operation is one of the text mining algorithms.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 513–519, 2005.

Section 2 describes proposed data mining approach that uses multiagent techniques, and our proposal approach is applied to data mining from document databases. Chapter 3 describes the Web information integration system with multiagent data mining. Section 4 describes conclusion and enhancing in a future.

2 Multiagent Data Mining with Databases

In KES2003, the multiagent technology is defined as a technology that processed information by cooperatively operating two or more independent programs (agent). [1]

Generally, multiagent technology is discuss with an autonomous control of an individual agent, but in this paper, we do not discuss it mainly.

A communication between agents between one to one, one to multi, multi to multi. In this paper, we use one to one communication by UNIX process communication, one to multi by Black board model.

2.1 Agent Definitions

The definition of agent which is used for data mining in this paper is defined as follows.

Query agent: Query agent receives used the database and the data mining algorithm from a user, and generates other agents. Query agent is generated at each demand of a user.

Mining agent: Mining agent generates DB-access agent, acquires data from DB-access agent, and applies data mining algorithm. Mining agent is generated of each applied mining algorithm.

DB-access agent: DB-access agent acquires data from the database, and sends it to mining agent. DB-access agent is generated of each database and of each mining agent.

Result agent: Result agent observes a movement of mining agents, and obtains result from mining agents. When result agent obtains all results, result agent arrangement/integrates, and shows it to a user.

Black board(BB): Place where results from data mining agent is written.

2.2 Flow of System

A flow of proposed system is defined as follows. (Fig. 1 shows flowchart of proposed system.)

1. A user generates Query agent, with setting the used database and the used data mining algorithm as its parameter.
2. The place of black board(BB) is set with Query agent.
3. Query agent generates Mining agent, and the place of BB is transmitted.
4. Query agent generates Result agent, and the place of BB is transmitted.

5. DB-access agent is generated, and Mining agent is accessed to the database.
6. DB-access agent gets data from the database.
7. Mining agent receives data from DB-access agent, and applies the data mining algorithm.
8. Mining agent writes the result of data mining on BB.
9. Result agent checks BB, and if all results are written, arranges the results and presents to the user.
10. All agents are eliminated.

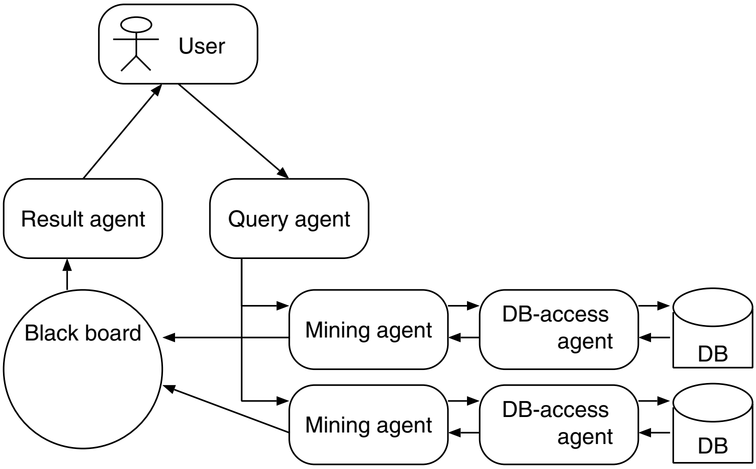


Fig. 1. Flowchart of Proposed System

2.3 Feature of Proposed Method

The proposal method has the following features.

The result of data mining can be made for more meaning result by building in the thesaurus agent as Mining agent, and making it can access the thesaurus database.

Query agent generates two or more Mining agent, it becomes possible to execute data mining algorithms in parallel. Moreover, it becomes possible that constructing the system and the access to the database and the processing of data are divided by separating DB-access agent accessed the database with Mining agent that processes data.

It becomes possible that the processing of each data mining algorithm and its arrangement/integration are separately thought by setting the agent which arranges the result. Moreover, it becomes easy to build arrangement/integration according to user's purpose into the system.

The system user comes to be able to construct the system corresponding to the purpose by recycling DB Agent and Mining Agent, and do tuning of Query agent and Result agent.

In this paper, the black board model with the file was handled with the interprocess communication on UNIX, but it can be easily enhanced to the communication on TCP/IP. Then, it is possible to enhance proposed approach to application to database that has been distributed on Internet. The problem of proposed approach is not using interprocess communication on UNIX but using black board model. Writing in the black board becomes a problem when the number of databases and data mining algorithm used increase, then the entire operation is influenced from the operation of the slowest agent. Therefore, the access to database and the processing of the data mining algorithm can be run parallel, but processing stops when checking results in the blackboard. It is necessary to consider that the maximum time is set to the black board writing check, and the system can show the result after each agent process.

3 Construction of Experimental Environment

We constructed an experimental environment which has multiagents with data mining algorithms to verify our proposed approach.

In this paper, proposed data mining using multiagent was applied to information integration system on Web, and the effectiveness was verified. In the proposed method, the part of the database access agent was changed to the Web access agent. Also, mining agent was changed to the information extraction agent from the HTML file. The information integration on the Web page can be thought just like the information integration from the database. Similarly, it can be thought that the information extraction operation is one of the text mining algorithms.

The constructed experimental environment was following.

We proposed following operation in the system that constructs with multiagent (Fig. 2). In this system, company information can be obtained by inputting URL of the company that wants to examine it is in the Web site that the user specified. The system has two main part of system.

One system is the system that retrieves company information by user's input and extracts information, and another one is a system that integrates company informations.

The system works by the flow from the following 1 to 5. The location of each operation is as shown in Fig. 2.

1. Read URL that the user input, and the Web page is preserved.
2. Information is extracted on the preserved Web page, and it preserves it in the XML file.
3. Two or more extracted XML files are integrated into one XML file.
4. The index is calculated from the integrated XML file, and added to the integrated XML file.
5. The result from XML file is displayed by Web a browser.

Essential information, the financial situation, employment information, and the index of the company are displayed as a result of this system. It is thought

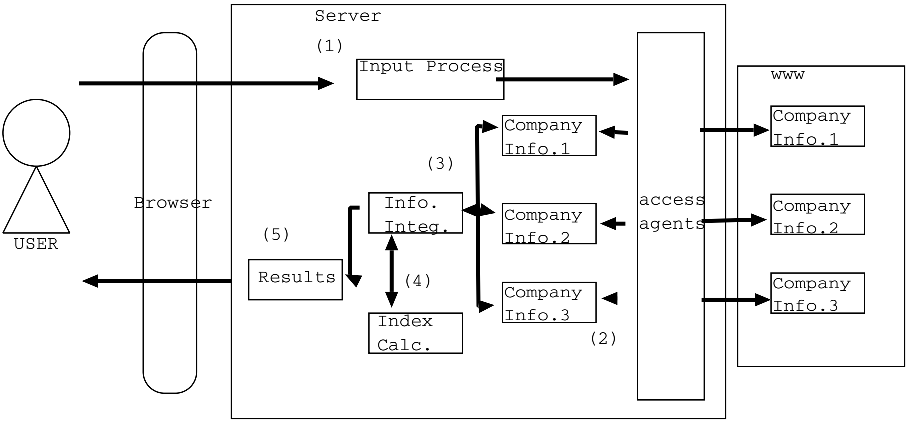


Fig. 2. Web Integration System

that more detailed information can be obtained by using not only basic information but also the index that can judge the company in the third person about the company.

3.1 Information Retrieve and Extract

In this system, at first, input information from Web browser, and the site on Web is preserved, and the system retrieves, and extracts necessary information from the preserved HTML file. In this operation, the text part of the corporate information is extracted from the tag of HTML by using the class of the pattern match of Java. Information on the extracted each item is put in the tag of specified XML. The XML file of each referred site is made by this operation.

3.2 Information Integration

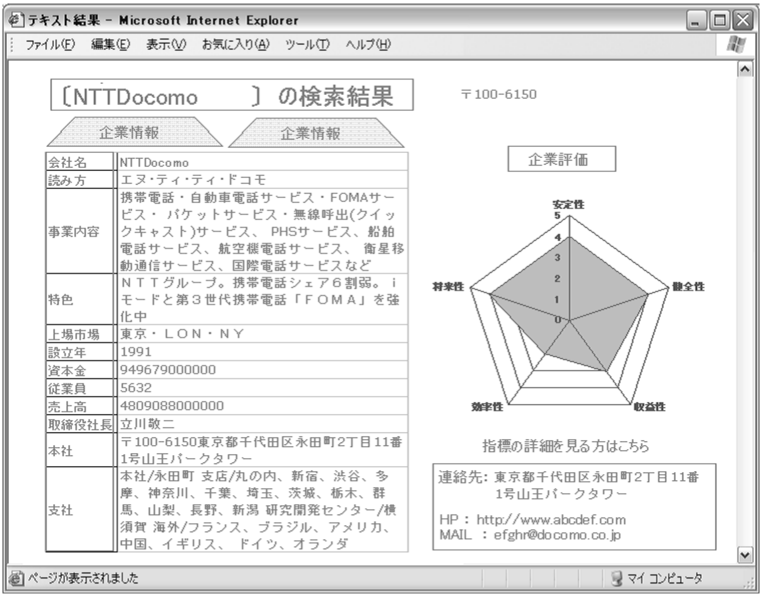
In the information integration, necessary information is extracted from the Web site by using the Java program, and each Web site is brought together in one XML file. In this research, it thought information was extracted from the Web site of various forms, and XML that was able to correspond to a lot of file formats was used. The XML file that extracts information on each tag in XML by using the Java program when information is integrated, extracts information from each Web site, and makes it individually is brought together in one XML file as information on one company.

3.3 Index Calculation

This system evaluates the company that uses the index as a material judged from a position the third person to know more detailed information about the corporate information. The index for the valuation of business enterprise is calculated by using extracted information, and it adds it to the XML file of the corporate information.

3.4 Show Results

The XML file that matches the index calculation result of making from the corporate information and such an index calculation is converted into the HTML form, and the corporate information is displayed. At this time, information such as the content of the work of the enterprise, the salaries, and branch offices is displayed in the form of the text besides the calculated index.



〔NTTDocomo 〕の検索結果

〒100-6150

会社名	NTTDocomo
読み方	エヌ・ティ・ティ・ドコモ
事業内容	携帯電話・自動車電話サービス・FOMAサービス・ パケットサービス・無線呼出(クイックキャスト)サービス、 PHSサービス、船舶電話サービス、航空機電話サービス、 衛星移動通信サービス、国際電話サービスなど
特色	ＮＴＴグループ。携帯電話シェア６割弱。ｉモードと第３世代携帯電話「ＦＯＭＡ」を強化中
上場市場	東京・ＬＯＮ・ＮＹ
設立年	1991
資本金	949679000000
従業員	5632
売上高	4809088000000
取締役社長	立川敬二
本社	〒100-6150東京都千代田区永田町2丁目11番1号山王パークタワー
支社	本社/永田町 支店/丸の内、新宿、渋谷、多摩、神奈川、千葉、埼玉、茨城、栃木、群馬、山梨、長野、新潟 研究開発センター/横須賀 海外/フランス、ブラジル、アメリカ、中国、イギリス、 ドイツ、オランダ

指標の詳細を見る方はこちら

連絡先: 東京都千代田区永田町2丁目11番
1号山王パークタワー

HP : http://www.abcdef.com
MAIL : efghr@docomo.co.jp

Fig. 3. Show Results

Information that cannot be used for a corporate name and making to the index in year of establishment etc. is displayed in the form of the text as shown in Fig. 3. Moreover, the index is shown in figure and the table like Star-chart for instance.

3.5 Experimental Results

In this paper, We confirm the operation of the information integration, and the index calculation in the system were done and verify the operation proof as the system that offered the valuation of business enterprise of the information extraction, Three kinds of experiments of (1) information extraction (2) information integration (3) index calculation and the result display were done in confirming the operation in each part. The program operated in each experiment without trouble as the result.

4 Conclusion

In KES2003 and KES2004, we proposed a technique for using multiagent technology in data mining intended for two or more text databases.

In this paper, proposed data mining using multiagent was applied to information integration system on Web, and the effectiveness was verified. In the proposed method, the part of the database access agent was changed to the Web access agent. Also, mining agent was changed to the information extraction agent from the HTML file. The information integration on the Web page can be thought just like the information integration from the database. Similarly, it can be thought that the information extraction operation is one of the text mining algorithms.

We constructed distributed Web integration with multiagent data mining for company information integration, and We verified its system. Three kinds of experiments of (1) information extraction (2) information integration (3) index calculation and the result display were done in confirming the operation in each part. The program operated in each experiment without trouble as the result.

There is XBRL for sharing the corporate information. [3] We want to examine integration with such data format in the future.

References

1. Niimi, A., Konishi, O.: Data Mining for Distributed Databases with Multiagents. KES'2003, Proceedings, PartII, Lecture Notes in Artificial Intelligence 2774, Springer:pp.1412–1418 (2003)
2. Niimi, A., Konishi, O.: Extension of Multiagent Data Mining for Distributed Databases. KES'2004, Proceedings, PartIII, Lecture Notes in Artificial Intelligence 3215, Springer:pp.780–787 (2004)
3. XBRL Japan, http://www.xbrl-jp.org/ (In Japanese)

Self-adjusting Programming Training Support System Using Genetic Algorithm

Eiji Nunohiro, Kenneth J. Mackin,
Masanori Ohshiro, and Kazuko Yamasaki

Department of Information Systems
Tokyo University of Information Sciences
1200-2 Yatoh-cho, Wakaba-ku, Chiba 265-8501, Japan
{nunohiro,mackin,ohshiro,yamasaki}@rsch.tuis.ac.jp

Abstract. Computer aided training has become an important method for improving computer education. For this research, we propose a programming training support system which targets understanding program structures which satisfy required program specifications. In our proposed training system, a given source code is broken up into separate puzzle pieces, and the user must layout the pieces in the correct order to reconstruct the program. The proposed system applies genetic algorithm (GA) and allows the system to self-adjust the difficulty of the programming problems matching the trainee's competency. We created a prototype system and applied it in a 1st year university programming course.

1 Introduction

Computer education, such as software operation and programming training, has received much interest recently in a wide range of academic programs, including high schools and liberal arts undergraduate programs. Computer aided training has become an important method for improving computer education. In the case of early programming training, there are many topics that need to be studied together, such as programming language syntax, programming design methods for example object oriented design, and problem solving algorithms. By applying an effective training program, the learning curve for students can be improved.

It is common to find an emphasis on language syntax in the early stages of programming training. But the essence of programming education is not mastering the language syntax, but understanding programming algorithms. For this research, we propose a programming training support system which targets understanding program structures which satisfy required program specifications. In our proposed training system, a given source code is broken up into separate puzzle pieces, and the user must layout the pieces in the correct order to reconstruct the program. The proposed system applies genetic algorithm (GA) and allows the system to self-adjust the difficulty of the programming problems matching the trainee's competency. We tested the proposed system on a prototype system and applied it to a 1st year university programming course.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 520–525, 2005.

2 Programming Problem Creation

There are many different methods for early stage programming training, for example:
method 1) the specification (problem) is supplied and a program which satisfies the specification (solves the problem) must be created;
method 2) the specification and partially completed program is supplied, and the program must be correctly completed.

For this research, we follow method 2 and propose a training system which takes into account the proficiency or level of the user. Here we emphasize 1) program flow comprehension, 2) program structure comprehension, and 3) program algorithm comprehension, rather than emphasizing memorization of programming language syntax.

Programming training from the viewpoint of program flow comprehension and program structure comprehension can be compared to solving a 2-dimensional puzzle by logically laying out the puzzle pieces. One must comprehend the general structure of the puzzle before laying out the individual pieces. From this standpoint, we propose a training system in which the system breaks up a program into puzzle pieces, and the user must reconstruct the program by selecting the correct program puzzle pieces in the correct order. In order for the system to break up a program into effective puzzle pieces, the system must select how many, and where to set the partition-points to divide the program. For this research, we applied genetic algorithm (GA) search to allow the system to self-organize the problem and select the optimal number of pieces and location of partition-points, depending on the proficiency of the user.

3 Program Puzzle Creation

Depending on the difficulty of the algorithm used in the original programming problem, and the difficulty of the created puzzle pieces, there is a possibility that the created problem becomes too easy or too difficult for the user and fails to improve the user's programming skills, causing an adverse affect to the user's motivation. For this reason it is important that to design the puzzle pieces considering the user's progress. We considered the following automatic puzzle creation algorithm.

Step 1) Analyze the program source code by determining for each program statement, a) the control structure depth (control information) and b) variable reference (reference information).

Step 2) Calculate the difficulty level for each statement using the above control and reference information.

Step 3) Apply genetic algorithm (GA) search to find the optimal combination of partition-points which best match the progress level of the user.

In the following sections we describe the genetic algorithm search applied.

3.1 Source Code Analysis

Each line of code is analyzed for data to be used in determining the difficulty of the partition. For each line of code, the control depth, and variable reference is calculated to measure the complexity of the source code.

Step a) control information analysis
The "control depth" for each statement is calculated. The control depth is the depth of the nest of control statements, such as if, for, and while statements.

Step b) reference information analysis
The "variable reference" count for each statement is calculated. The following criteria are used to determine the variable reference for statement S:

- a variable is set in statement S
- a variable is referenced in statement S

The variable reference count for a statement is the total number of times a variable is set or referenced in the statement.

An example of source code analysis is shown in Figure 1.

```
      Program                                              Depth     Reference
 1:   public class EvenOrOdd {                               0           0
 2:     public static void main(String[] args) {             1        1(Def.1)
 3:       int Evev = 0;                                      1        1(Def.1)
 4:       int Odd = 0;                                       1        1(Def.1)
 5:       for(int i=0 ; i<=100 ; i++){                       1     4(Def.2, Set2)
 6:         if(i % 2 == 0){                                  2        1(Set1)
 7:           Even = Evev+1;                                 3     2(Def.1, Set1)
 8:         }else{                                           2           0
 9:           Odd = Odd+1;                                   3     2(Def.1, Set1)
10:         }                                                2           0
11:       }                                                  1           0
12:       System.out.println(" Number of Even : "+Even);     1        1(Set1)
13:       System.out.println("Number of Odd"+Odd);           1        1(Set1)
14:     }                                                    0           0
15:   }                                                      0           0
```

Fig. 1. Example source code analysis

The sample program in Figure 1 counts the number of even and odd numbers from 1 to 100. The control depth and variable reference values are shown to the right of each statement.

3.2 Program Partition Points

The program puzzle pieces are created by selecting partition points between two lines of code. The difficulty of the puzzle is dependent on the number of pieces and the location of the partition points. If a partition occurs at a location with high control depth and high reference count, the puzzle can be assumed to have a high difficulty. From this assumption, we define the following function to evaluate the difficulty of the partition. Partition difficulty $_k$ is the evaluated difficulty when a partition occurs after program code statement k. Partition difficulty Pd $_k$ is defined as:

$$Pd_k = \text{control depth} * \text{reference count} \tag{1}$$

The difficulty of the puzzle is also affected by the difficulty of the program algorithm used in the given program source code, but this degree difficulty is dependent on the prior knowledge of the individual user. For this research we let the user decide on the degree of difficulty for an algorithm.

3.3 Program Selection by GA

The partition pattern (number of partitions and location of partitions) is selected using the user's progress level and partition difficulty for each possible partition point. We apply genetic algorithm (GA) to select the partition pattern. GA is a general purpose search algorithm, which is effective in searching a vast search space for an optimal solution when the solution is not known beforehand.

In GA, an initial set of random search points (population) is created, and each point is evaluated for the fitness or effectiveness using a fitness function. The points are then randomly modified using operations such as crossover and mutation. The process of evaluation and modification is repeated until an optimal solution is found. The GA procedure used in this research is described in the following.

Step 1. Creation of initial population
An initial population, where individual (or chromosome) expresses a specific partition pattern, is created randomly.

Step 2. Fitness evaluation
Each chromosome is evaluation with the fitness function.

Step 3. Solution check
If a predefined fitness or number of cycles is achieved, then exit the algorithm.

Step 4. Selection
The fitness values are used in a selection algorithm, where individuals are selected to the next generation. A combination of elite selection, which guarantees individuals with high fitness, and roulette selection, allows individuals with lower fitness to be selected, is used.

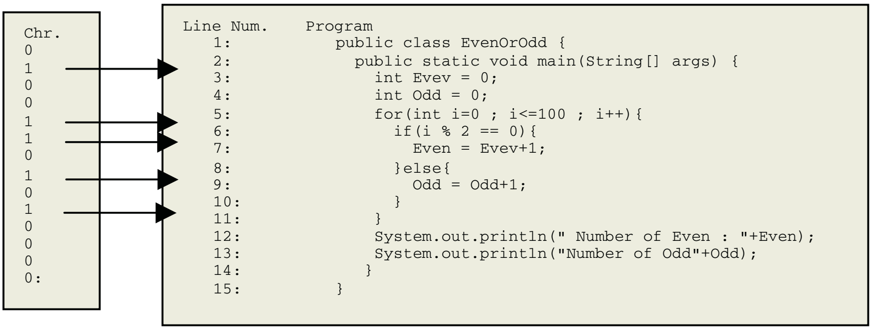
Step 5. Crossover
Using a crossover algorithm, a portion of two chromosomes are swapped. A single point crossover algorithm is used.

Step 6. Mutation
Using a mutation algorithm, a portion of a chromosome is modified.

Step 7. Repeat from step 2.

3.4 Chromosome Expression

The partition pattern expressed as a binary string is used as the chromosome in the genetic algorithm. The length of the chromosome is 1 less than the lines of code (statements) in the program. In the chromosome, the value 1 indicates a partition at that location (line), and 0 indicates no partition at that location. Figure 2 illustrates the relationship between chromosome and partition points. In figure 2 an example of a 15 line program is shown, and the partition points are the locations where the chromosome value is 1, i.e. after lines 2,5,6,8,10.



```
Chr.
0
1
0
0
1
1
0
1
0
1
0
0
0
0:
```

```
Line Num.    Program
   1:           public class EvenOrOdd {
   2:             public static void main(String[] args) {
   3:               int Evev = 0;
   4:               int Odd = 0;
   5:               for(int i=0 ; i<=100 ; i++){
   6:                 if(i % 2 == 0){
   7:                   Even = Evev+1;
   8:                 }else{
   9:                   Odd = Odd+1;
  10:                 }
  11:               }
  12:               System.out.println(" Number of Even : "+Even);
  13:               System.out.println("Number of Odd"+Odd);
  14:             }
  15:         }
```

Fig. 2. The relationship between chromosome and partition points

3.5 Fitness Evaluation

Fitness is evaluated using the user's level of progress, target problem difficulty, and difficulty of the created program puzzle. Below we describe the fitness evaluation method.

User's Level of Progress

The user's level of progress is defined as the puzzle difficulty of the last cleared problem.

Puzzle Difficulty

The puzzle difficulty is the sum of the partition difficulty value for the given program puzzle chromosome, calculated with the following eqation.

$$\text{puzzle difficlty} = \sum_{k=1}^{n} \text{Pd}_{\text{k}} * \text{Chr}_{\text{k}} \tag{2}$$

where:

n is number of lines of code - 1, Pd_{k} is the puzzle difficulty at position k, Chr_{k} is the value of the chromosome at position k .

Fitness Function

The fitness function is defined as the following

$$\text{fitness(p,d,s)} = (\text{p} * \text{s}) - \text{d} \tag{3}$$

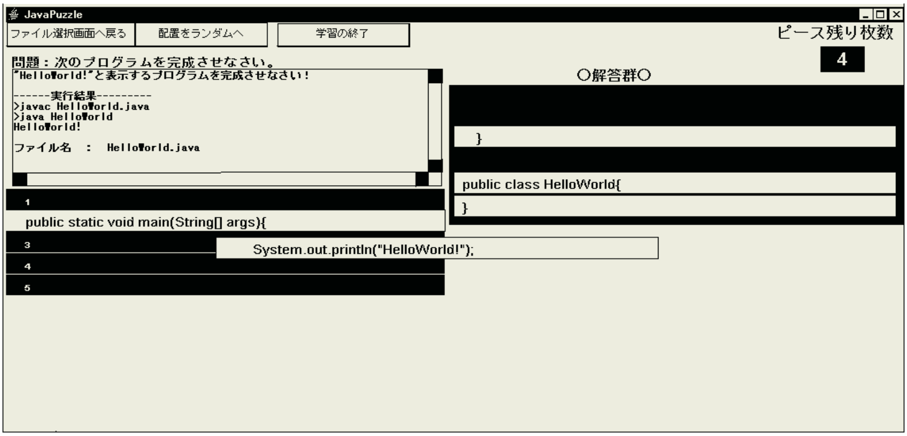
where :

p is the progress level, d is the puzzle difficulty, s is the expected degree of difficulty of the created problem (e.g. s = 1 for same difficulty, s = 2 for double difficulty).

The chromosome with the fitness value closest to 0 is selected as the optimal program puzzle combination for the user.

4 Application Results

The proposed programming training support system was applied in a 1st year Java programming course at Tokyo University of Information Sciences. The proposed method is essentially language independent, and if the source code analysis section is modified, the system could be applied to other structural programming languages such as C. Figure 3 shows the program puzzle solving stage of the system.



問題：次のプログラムを完成させなさい。

```
"HelloWorld!"と表示するプログラムを完成させなさい！

------実行結果---------
>javac HelloWorld.java
>java HelloWorld
HelloWorld!

ファイル名  :  HelloWorld.java
```

```
public static void main(String[] args){
System.out.println("HelloWorld!");
}
public class HelloWorld{
}
```

Fig. 3. Application screenshot of proposed system

5 Conclusion

One of the most important points to carry out an effective programming course, is how to attend to the different skills and levels of each student. For larger classes, it becomes very difficult to teach each of the students individually. The proposed self-adjusting programming training support system can provide tailored programming exercises for students of different achievement levels, and efficiently support the learning environment. The self-organizing nature of the proposed method allows the system to create many different patterns of problems from a limited number of sample source code, and the genetic algorithm search enables the system to select different problems for repeated trials, and effectively supports the programming training of the user. The proposed training support system is still in the prototype phase, but we plan to further improve the quality of the system, including reevaluation of the genetic algorithm used.

References

1. Jain, L.C., et.al. (eds): Virtual Environments for Teaching and Learning, World Scientific (2002)

Waste Incinerator Emission Prediction Using Probabilistically Optimal Ensemble of Multi-agents

Daisuke Yamaguchi[1], Kenneth J. Mackin[2], and Eiichiro Tazaki[1]

[1] Department of Electronics and Information Engineering, Toin University of Yokohama
1614 Kurogane-cho, Aoba-ku, Yokohama 225-8502, Japan
{yamaguti,tazaki}@intlab.toin.ac.jp
[2] Department of Information Systems, Tokyo University of Information Sciences
1200-2 Yatoh-cho, Wakaba-ku, Chiba, Japan
mackin@rsch.tuis.ac.jp

Abstract. The emission of dioxins from waste incinerators is one of the most important environmental problems today. It is known that optimization of waste incinerator controllers is a very difficult problem due to the complex nature of the dynamic environment within the incinerator. In this paper, we propose applying a probabilistically optimal ensemble technique, based on fault masking among individual classifier for N-version programming. We create an optimal ensemble of neural network trained multi-agents and use the majority voting result to predict waste incinerator emission. We show that an optimal ensemble of multi-agents greatly improves the prediction error rate of emission of dioxins.

1 Introduction

Dioxin emission from waste incinerator plants is one of the hottest ecological problems today. In waste incinerator plants, the chemical reactions in the incinerator occur under a very dynamic environment, making its control a very complex task, and current state-of-the-art incinerator facilities have not succeeded in completely removing the dioxin emission. The volume, density and contents of the garbage to be incinerated are not constant, so it is impossible to control the combustion as in a laboratory environment. One of the causes of dioxin emission in waste incinerator plants is due to the fluctuations in the amount of garbage fed in the incinerator. The fluctuation in garbage fed leads to temporary deterioration of the combustion state (i.e. oxygen rate), and short peaks of dioxin emission occur.

There has been past research in intelligent estimation of dioxin emission from waste incinerators. Fujiyoshi et al. [1] has proposed applying fuzzy control to incinerator control to decrease the dioxin emission. Ichihashi et al. [2] has applied statistical analysis to calculate the correlation of various input signals with dioxin emissions. Fukushima [3] has proposed applying fractal fuzzy control in order to estimate and control dioxin emission.

For this research, we investigate the validity of applying neural network trained multi-agents for the prediction of dioxin emission to be used for the combustion control in order to decrease the dioxin emission. Neural networks, as with other training based classifiers, inherently have a risk that when classifying an untrained data set,

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 526–532, 2005.

the classifying error rate may be much worse than the training result. In order to overcome this risk, we applied a probabilistically optimal ensemble technique proposed by Imamura et al. [4] to N-version programming of software agents. Our purpose is to research the validity of using a training based classifier method for emission prediction, compared against the results of previous statistical methods and fuzzy decision methods. We plan to continue this approach and apply the proposed method to actual incinerator plant controllers to test the validity of the method.

2 Requirements for Waste Incinerators

For this research, we used real waste incinerator data provided by Hitachi Zosen Corporation. Fluidized bed incinerator data from the Ryotsu City Clean Center in Niigata prefecture, Japan, was used.

The data consists of the following sensor values measuring various conditions of the incinerator. Flapper angle(0 - 100.00%), oxygen concentration in incinerator exit (0 - 25.000%), garbage rate(t/H), incinerator temperature(0 - 1200.0%), carbon monoxide concentration(0 - 500.0ppm), incinerator pressure(-2000.0 - 1000.0ppm), cooling liquid rate(0 - 1.0000m3/h), conveyer belt speed(0 - 7.000rpm), primary air supply(0 - 7.500KNm3/h), secondary air supply base(0 - 7.500Nm3/h), secondary air supply modification(0 - 7.500KNm3/h).

The flapper is lifted as garbage is carried by the conveyer belt, and the flapper angle is used as the measure of garbage volume. The above sensor data was collected in approximately 2 second intervals.

It is known that CO (carbon monoxide) concentration over 100ppm show strong correlation with dioxin concentration. For this research, we use the CO concentration as the target output, and aim to reduce the average CO concentration as well as to reduce the number of CO concentration peaks over 100ppm.

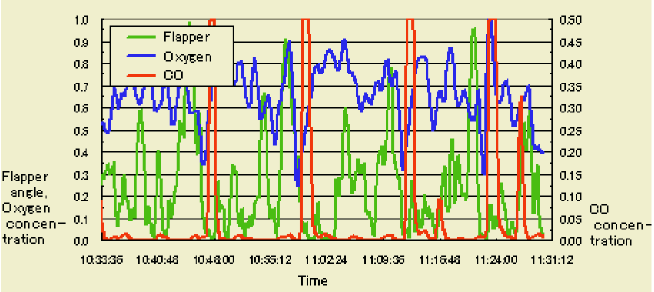


Fig. 1. Relationship of flapper angle, oxygen concentration, and CO concentration

Figure 1 shows a portion of the collected time series data. From the collected data, it can be observed that when the flapper angle (garbage volume) increases, after a time delay the oxygen concentration decreases, and after further time delay the carbon monoxide (CO) concentration increases. This can be explained by the following. The increased garbage measured by the flapper takes some time before arriving at the incinerator. The increased garbage in the incinerator increases the combustion and consumes more oxygen, which lowers the oxygen concentration. The temporarily decreased oxygen concentration deteriorates the combustion state and a peak in CO occurs due to imperfect combustion. Since the CO sensor is placed at the incinerator emission, the peak in carbon monoxide concentration is displayed after a further time delay.

From the above observation, for this paper we especially concentrate on flapper angle and oxygen concentration as input for predicting CO output.

3 Incinerator Control Using Multi-agent System

For each of the different types of incinerator sensor data, there is an apparent correlation just described, but direct correlation between the sensor data and carbon monoxide concentration is not very strong. This is because the environment in the incinerator is a complex dynamic environment in which the different items are dependent on each other, and is not a simple dependency relationship.

Artificial neural networks (ANN) can be characterized by its "black box" approach to learn and classify complex data patterns. For this research, we propose applying 3 layer network structure (1 input layer, 1 hidden layer, 1 output layer) for the training of an incinerator emission prediction agent, using the neural network to learn the complex relationship between incinerator sensor data.

The proposed multi-agent system for incinerator emission prediction is part of a larger incinerator controller system plan. The incinerator controller system will consist of two separate multi-agent systems, the dioxin prediction section and the combustion controller section. Each section will use a separate, independently trained multi-agent system. The dioxin prediction multi agent-system uses incinerator sensor input and predicts the carbon monoxide (hence dioxin) emission rate before the actual emission occurs. The combustion controller multi-agent system uses the output from the dioxin prediction system as a trigger, as well as incinerator sensors for input, and outputs changes in incinerator control values in order to decrease the predicted carbon monoxide emission.

For this paper, we will propose methods applying N-version programming using software agents to construct the dioxin prediction section. We will discuss the combustion controller system in future works.

4 Dioxin Prediction with Software Agent

First we describe the basic software agent for dioxin prediction. We considered the 3 layer artificial neural network (1 input layer, 1 hidden layer, 1 output layer) as the basic training classifier in the agent. We use a sigmoid function for the synapse function of the neuron, with back propagation (BP) training of the incinerator data. The number of hidden neurons was decided by results of preliminary experiments of the neural network.

As a preliminary experiment, we tested a neural network which took all of the sensor data except carbon monoxide concentration values for input, and the single output of the network was used to predict the correct carbon monoxide concentration. Time delay of input data was not considered here.

For the network training we used the database of collected incinerator sensor data, and applied BP training based on the difference between predicted carbon monoxide concentration and the actual carbon monoxide concentration recorded for the same time frame.

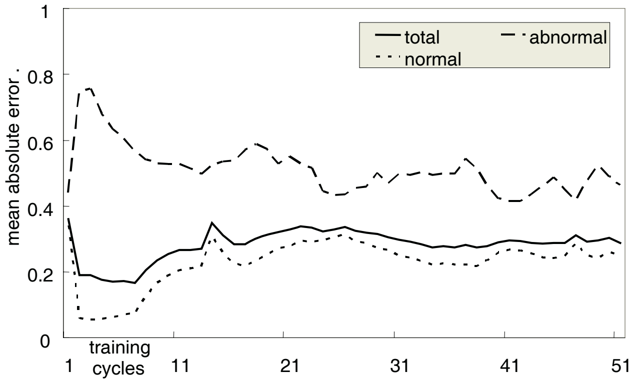


Fig. 2. Preliminary experiment results of prediction error for untrained data

Figure 2 shows the training results of the preliminary experiment which predicted the carbon monoxide values directly using the neural network. From the results of the preliminary experiment, we found that the prediction accuracy is completely different between normal range carbon monoxide values, and high carbon monoxide values. The network learned to accurately predict normal range carbon monoxide values fairly quickly, but the same network failed to learn abnormal (high) range carbon monoxide values during the same training period. When network training was continued in order to increase the abnormal range carbon monoxide prediction, this time the accuracy of normal range carbon monoxide prediction deteriorated. This finding confirmed our initial estimate that it would be difficult to train the neural network due to the complexity of the correlation between carbon monoxide concentration and each of the other sensor values.

For this reason, we decided to focus on detection of abnormally high carbon monoxide emission (>100ppm) as the preliminary goal of the dioxin prediction network. The network output was changed from direct carbon monoxide concentration prediction value, to binary output where 1 predicts high carbon monoxide concentration (>100ppm) and 0 predicts normal carbon monoxide concentration (<= 100ppm).

As for the neural network input, we considered the possibility that the large number of input nodes increases the problem domain and complicates the classification, causing an adverse affect on the network training efficiency. With this assumption, we decided to minimize the number of input nodes in order to first achieve a workable learning curve and prediction accuracy.

It can be assumed that there must be some relationship between oxygen concentration, flapper angle and carbon monoxide concentration, from the similar changes seen in time-series data as in Figure 2. Based on this assumption, for the initial model we use only flapper angle and oxygen concentration data as neural network input. Further, we can see that flapper angle, oxygen concentration and carbon monoxide concentration each show a particular time delay in their relationship. For this reason, in order to predict the carbon monoxide value for a given instance, time delay for the flapper angle and oxygen values must be taken into account. Data at some fixed time frame previous to the given output instance should be used as the input data. Recurrent network structures could be used to automatically treat such time sequence data effectively, but for the initial model, we map sequential data of flapper angle and oxygen concentration of specified time delay to individual input nodes to the network. Specifically, we used 60 second delay for flapper angle (t-60) and 30 second delay for oxygen data (t-30), to predict the emission for time t.

5 Experiment Results of Single Agent

We trained the software agent with the above described neural network using BP and evaluated the prediction accuracy. A standard sigmoid function was used as the neuron's base synapse function. The number of neurons used in each layer was 3 input neurons (1 flapper input, 1 oxygen input and 1 fixed input), 6 hidden layer neurons, and 1 output neuron.

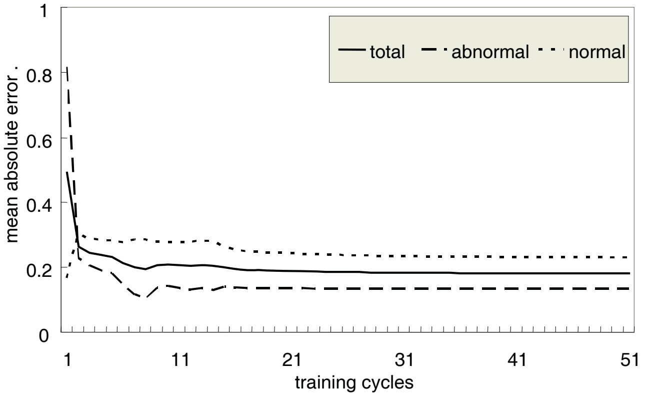


Fig. 3. Prediction error of untrained data using single neural network

For the training data, 100 cases of normal range carbon monoxide data and 100 cases of abnormal (>100ppm) carbon monoxide data, for a total of 200 cases were randomly selected from the incinerator sensor database. For the untrained data used to plot the training curve of network accuracy, 100 cases of normal range carbon monoxide data and 100 cases of abnormal range (>100ppm) carbon monoxide data, for a total of 200 cases were randomly selected from the incinerator sensor database.

Figure 3 shows the change in output error for the untrained dataset of the proposed neural network. The absolute output error for normal range carbon monoxide values, absolute output error for abnormal range (>100ppm) carbon monoxide values, and absolute output error for all values is graphed.

The absolute output error for a single neural network shown in Figure 4 was 0.18 for all values, 0.23 for normal range carbon monoxide values, and 0.13 for abnormal range carbon monoxide values. As was seen in the preliminary experiment, as the network trained to decrease the abnormal range output error, the normal range output error in turn became higher.

6 Probabilistically Optimal Ensemble of Multi-agents

Neural networks, as with other training based classifiers, inherently have a risk that when classifying an untrained data set, the classifying error rate may be much larger than the training result. In order to overcome this risk, we applied a probabilistically optimal ensemble technique proposed by Imamura et al. [4] to N-version programming of software agents.

Fault masking in N-version programming assumes that the individual members give completely independent results. If certain members give similar output, then correct fault masking will not occur. In a probabilistically optimal ensemble, the members of the ensemble are chosen so that the members are correctly independent of each other. This is realized by selecting members so that the measures error rate of the ensemble comes closest to the expected error rate of the ensemble. If the members are correctly independent of each other, then proper fault masking should allow the measured error rate to become very close to the expected error rate.

The expected failure rate f of the probabilistically optimal ensemble can be calculated by the following equation [4]

$$f = \sum_{k=m}^{n} \binom{n}{k} (1-p)^{n-k} p^{k} \tag{1}$$

where p is the failure rate of each individual, n is the size of the ensemble, m is the minimum number of faulty outputs for an ensemble to fail. We assume the same failure rate p for individuals for simplicity.

In the case where the ensemble has 3 members, majority vote (2 votes) for output, and $p = 0.19$, then $f = 0.086$.

For our research we trained 6 neural network agents using different initial weights, and the same training set. Using the same training set, we compared the measured ensemble failure rate for ensemble size 3, for all combinations of agents. Using individual failure rate $\mathrm{p} = 0.19$, the expected ensemble error rate was $f = 0.086$.

Table 1 shows the results of the measured ensemble failure rate. The ensemble with the closest ensemble failure rate (0.085) was selected as the optimal ensemble.

The selected optimal ensemble was evaluated using untrained data. The resulting ensemble error rate was $f = 0.087$, indeed very close the training ensemble error rate, and vastly improved over the error rate 0.19 for the single neural network agent.

7 Conclusion

In this research we applied a probabilistically optimal ensemble technique [4] to create an N-version programming classifier system using software agents trained by 3-layer neural networks to predict incinerator emission. We were able to confirm that by using an optimal ensemble of independent software agents, the classification error rate can be greatly reduced.

Table 1. Measured ensemble failure rate

ANN no.	ensemble failure rate
0,1,2	0.11
0,1,3	0.115
0,1,4	0.095
0,1,5	0.085
0,2,3	0.115
0,2,4	0.12
0,2,5	0.1
0,3,4	0.13
0,3,5	0.12
1,2,3	0.13
1,2,4	0.135
1,2,5	0.115
1,3,4	0.14
1,3,5	0.13
1,4,5	0.115
2,3,4	0.145
2,3,5	0.135
2,4,5	0.125

For future works, we will consider methods to improve prediction accuracy of the individual neural network, including the increase in the types of sensor input data, reevaluation of neural network structure, combining fuzzy rules to treat input data, as well as effect of using different base synapse functions for neurons.

References

1. Makoto Fujiyoshi, Ryutaro Fukushima, Mitiharu Masuya, “Intelligent Control System for Fluidized Bed Incinerator”, Proceedings of 18th Fuzzy System Symposium, Japan, pp.25-28, 2002
2. H. Ichihashi, et. al, “Fuzzy Bi-plot of Correlation Analysis for Waste Incinerator”, Proceedings of 19th Fuzzy System Symposium, Japan, 2003
3. R. Fukushima, “Fractal Fuzzy Intelligent Control System”, Proceedings of 20th Fuzzy System Symposium, Japan, 2004
4. Kosuke Imamura, Kris Smith, “A Probabilistically Optimal Ensemble Technique for Training Based Classifiers”, Proceedings of Joint 2nd International Conference on Soft Computing and Intelligent Systems and 5th International Symposium on Advanced Intelligent Systems, Japan, 2004

Cooperative Control Based on Reaction-Diffusion Equation for Surveillance System

Atsushi Yoshida, Katsuji Aoki, and Shoichi Araki

Advanced Technology Research Laboratories, Matsushita Electric Industrial Co., Ltd,
3-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
{yoshida.a2c,aoki.katsuji,araki.shoichi}@jp.panasonic.com

Abstract. This paper proposes a new cooperative control model for the surveillance system which consists of plural Pan-Tilt-Zoom cameras and no central control unit. Each camera adjusts their observation area to decrease blind spots in the whole surveillance area by the control model based on reaction-diffusion equation. Simulation results have shown that the surveillance system is able to cover new blind spots when some cameras are removed or are rearranged in their placement.

1 Introduction

Many theoretical models based on reaction-diffusion phenomena have been proposed to account for patterning phenomena in morphogenesis [1,2,3,4]. In the engineering field, "reaction-diffusion equation on a graph [5,6]" is one of the promising approaches for designing a self-organizing system which is adaptive, scalable and fault-tolerant. The methodology has been applied to various problems in the design of autonomous decentralized systems such as traffic signal control and cooperative exploration by mobile robots[7,8].

In this paper, we propose a new cooperative control model for surveillance system. The system consists of plural Pan-Tilt-Zoom cameras. Each camera's observation area is autonomously arranged in response to the conditions of the surveillance area and changes in the system structure by using a spatial pattern generator based on a reaction-diffusion equation. Fault-tolerance is one of the significant requirements for a surveillance system. In the proposed system, for example, the plural cameras autonomously cooperate with one another to decrease blind spots in the whole surveillance area even if some cameras are removed from the system by some troubles. Each camera adjusts their observation area in order to overlap into the neighboring areas observed by other cameras or neighbor the boundary in the surveillance area. We present two simulation results which show the proposed system has adaptability to organize the arrangement pattern of each camera's observation area when some cameras are removed or are rearranged in their placement.

2 Outline of the Surveillance System

The surveillance system consists of plural Pan-Tilt-Zoom cameras and no central control unit. The system organizes the arrangement pattern of each camera's observa-

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 533–539, 2005.

tion area in order to decrease blind spots in the whole surveillance area. In the field of computer vision, a cooperative distributed vision system which consists of plural active cameras for continuous tracking of multi-targets in a wide-spread area has been proposed[9]. In contrast, the purpose of the proposed system is not to track objects but to maintain the observation of the whole surveillance area by plural cameras. Additionally, we aim that the system satisfies the following requirements.

1. Assign each camera's observation area to observe whole surveillance area without being dependent on the number and locations of cameras
2. Reconstruct arrangement pattern of each observation area in order to cover new blind spots when some cameras are removed or turned off due to trouble etc.

In order to satisfy the above-mentioned requirements, in the proposed system, each camera communicates mutually and repeats the following processing.

1. Exchange the observation area data which is calculated from pan θ_{Pan} , tilt θ_{Tilt} and focus ζ , with other cameras.
2. Detect neighboring areas observed by other cameras or neighboring boundaries in the surveillance area.
3. Adjust pan θ_{Pan} , tilt θ_{Tilt} and focus ζ , so that an observation area overlap into the neighboring areas observed by other cameras or neighbor the boundaries of the surveillance area.

In order to control the whole system by local adjustment of each camera, each ignorant of state of the whole system, the third processing step applies the control model based on "reaction-diffusion equation on a graph" to adjustment of each camera's pan θ_{Pan} , tilt θ_{Tilt} and focus ζ .

3 Design of Cooperative Control Model for Surveillance System

In this section, in order to describe our proposed model, first we introduce the idea of "reaction-diffusion equation on a graph". Secondly we explain the design of a cooperative control model for decreasing the blind spots in the surveillance area.

3.1 Reaction-Diffusion Equation on a Graph

Yuasa *et. al.* proposed to express a system that consists of homogeneous elements by a graph and to design the behaviors of individual elements in such a way that the whole system became desirable[5]. V denotes a set of vertices (subsystems), E denotes a set of edges (interaction), and $G = \{V, E\}$ denotes a finite graph with boundary as shown in Fig1. $C(V)$ and $C(E)$ denote the whole set of real functions on the V and E , respectively. For $f \in C(V)$, it is defined the differential operator d , called codifferentiation, as $d : C(V) \to C(E)$, $df(e) = f(t(e)) - f(o(e))$ where $o(e) \in V$ is the initial vertex of $e \in E$ and $t(e) \in V$ is the terminal vertex of $e \in E$, df is called the gradient of f . In order to evaluate each vertex $u \in V$ and edge e , we define $W_0(f) = \sum F_R(f(u))$, $W_1(f) = \sum F_D(d\mathbf{f}(u))$ where F_R and F_D are differentiable function, $d\mathbf{f}(u) = [df(e)]_{e \in E(u)}$. In order to evaluate the whole system, we define the function as

$$W(f) = W_0(f) + W_1(f) \tag{1}$$

Then we calculate the partial derivative of $W(f)$ with respect to f on each vertex, and we define the evolution equation of f as

$$\frac{\partial f}{\partial t} = -\frac{\partial W(f)}{\partial f} \tag{2}$$

Equation 2 calculates f which gives the optimal solutions of $W(f)$ in the function space. Then it is proved that Eq.2 expresses reaction-diffusion equation and the alternation of f is determined by only the state of subsystem which has f and interaction with its neighbors by Yuasa *et. al.* Thus, in order to control the whole system by local adjustment of each subsystem, we just have to design the $W_0(f)$ and $W_1(f)$ which express the desirable state of the system and make the evolution equation of f.

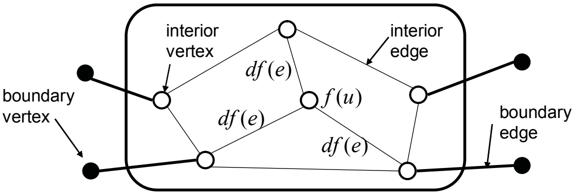


Fig. 1. Graph expression of system with boundary

3.2 Cooperative Control Model for Decreasing Blind Spots in the Surveillance Area

In order to organize the arrangement pattern of each camera's observation area which decreases blind spots in the whole surveillance area, we defined $W_0(f)$ and $W_1(f)$ as follows. First, each camera $i = 1, \cdots, N$ adjusts their observation area to a predetermined state. Thus we define

$$W_{0_i} = \alpha(\theta_{\text{Pan}} - \psi_{\text{Pan}})^2 + \alpha(\theta_{\text{Tilt}} - \psi_{\text{Tilt}})^2 + \beta(\zeta - \psi_\zeta)^2 \tag{3}$$

where ψ_{Pan}, ψ_{Tilt} and ψ_ζ are predetermined values. α and β are coefficients. Secondly, in order to decrease blind spots between observation area and neighboring area observed by other camera, each camera adjusts their observation area to overlap into neighboring areas. Moreover, each camera adjusts the size of their observation area to become the same size as neighboring areas. Thus we define the evaluation function for the relation of each observation areas as

$$W_{11i} = \gamma \sum_{j}^{n} (O_j(\theta_{\text{Pan}}, \theta_{\text{Tilt}}, \zeta) - O_{\text{target}})^2 + \delta \sum_{j}^{n} (S(\theta_{\text{Pan}}, \theta_{\text{Tilt}}, \zeta) - S_j)^2 \tag{4}$$

where $j = 1, \cdots, n$ is the number of neighboring areas. $O_i(\theta_{\text{Pan}}, \theta_{\text{Tilt}}, \zeta)$ is a function which calculates the depth of overlap with neighboring areas observed by j-th camera. O_{target} is target depth of overlap area. $S(\theta_{\text{Pan}}, \theta_{\text{Tilt}}, \zeta)$ is a function which calculates the size of observed area. S_j is the size of area observed by j-th camera. γ and δ are coefficients. Finally, in order to decrease blind spots between observation area and neighboring boundary, each camera adjusts their observation area to the neighboring boundary. Thus we define evaluation function of the relation between observation area and neighboring boundary as

$$W_{12i} = \varepsilon \sum_{k}^{m} (D_k(\theta_{\text{Pan}}, \theta_{\text{Tilt}}, \zeta))^2 \tag{5}$$

where $k = 1, \cdots, m$ is the number of neighboring boundaries. $D_k(\theta_{\text{Pan}}, \theta_{\text{Tilt}}, \zeta)$ is a function which calculates the distance to k-th neighboring boundary. ε is a coefficient. We define evaluation function for the whole surveillance system as

$$W(\theta_{\text{Pan}}, \theta_{\text{Tilt}}, \zeta) = \sum_{i}^{N} (W_{0i} + W_{11i} + W_{12i}) \tag{6}$$

We calculate the partial derivative of $W(\theta_{\text{Pan}}, \theta_{\text{Tilt}}, \zeta)$ with respect to pan θ_{Pan}, tilt θ_{Tilt} and focus ζ on each camera in order to decline $W(\theta_{\text{Pan}}, \theta_{\text{Tilt}}, \zeta)$ to the local minimum. Thus we define

$$\begin{gathered} \partial\theta_{\text{Pan}} / \partial t = -\partial W(\theta_{\text{Pan}}, \theta_{\text{Tilt}}, \zeta) / \partial\theta_{\text{Pan}} \\ \partial\theta_{\text{Tilt}} / \partial t = -\partial W(\theta_{\text{Pan}}, \theta_{\text{Tilt}}, \zeta) / \partial\theta_{\text{Tilt}} \\ \partial\zeta / \partial t = -\partial W(\theta_{\text{Pan}}, \theta_{\text{Tilt}}, \zeta) / \partial\zeta \end{gathered} \tag{7}$$

Equation7 is the evolution equation of each camera's pan θ_{Pan}, tilt θ_{Tilt} and focus ζ in order to decrease blind spots around their observation area. In the proposed model, each camera's pan θ_{Pan}, tilt θ_{Tilt} and focus ζ are controlled by Eq.7.

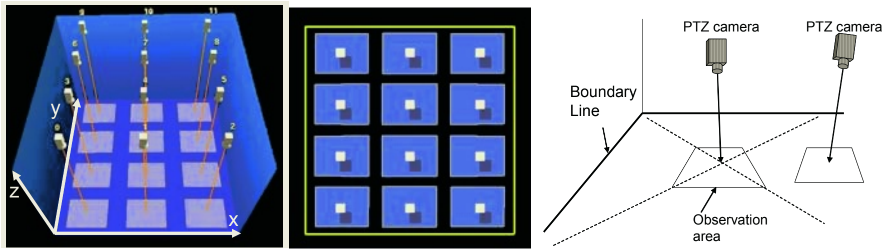
4 Simulation

In this section, we show two simulation results to investigate the following points.

(Simulation1) Adaptability to organize the arrangement pattern of each camera's observation area in order to decrease blind spots in the whole surveillance area without being dependent on the number and locations of cameras.

(Simulation2) Robustness of surveillance system in the case that some cameras are removed from the surveillance system.

In the simulation, as shown in Fig.2 (a), we define a rectangular space in which each camera is attached in the ceiling (xy plane where z = 2.5 [m]) turned towards the floor (xy plane where z = 0 [m]) defined as the surveillance area. In order to simplify, the shape of an observation area is approximated with the rectangle inscribed in an actual observation area as shown in Fig2(b). We use the real camera data (Panasonic WV-CS850A) as velocity and movable range of Pan, Tilt and Focus and size of CCD. Each camera communicates to exchange the data of their observation area with other cameras every 100 msec.



(a) (b) (c)

Fig. 2. (a)Surveillance space (width 3.6[m] * depth 3.6 [m]) and Pan-Tilt-Zoom cameras. (b) Arrangement pattern of areas observed by each camera that watched from a top of the surveillance space. (c) Detection of neighboring areas observed by other camera or boundary lines in the surveillance area

In order to detect neighbors, cameras divide the surveillance area into four domains by the straight line extended from the center of their observation area to each vertex as shown in Fig.2(c). Then, in each divided area, cameras choose the center of the area observed by other cameras which are the nearest to its center as the neighboring observation area. If there is no center of area observed by other cameras in divided area, cameras choose the longest boundary in divided area as boundary to neighbor. In order to calculate evolution Eq.7, we used 4th Runge-Kutta method. Additionally, in Eq.4, O_{target} is defined as 20% of the maximum depth of overlap into neighboring areas observed by other cameras. In order to evaluate the arrangement pattern of each camera's observation area, we defined cover rate, C, and redundant rate, R, as $C = OA/WA$, $R = 1-(OA/TA)$ where WA is the size of the whole surveillance area, OA is the size of logical sum of observation areas, and TS is the total size of all observation area.

Simulation 1

We show examples of an arrangement pattern of each camera's observation area which was organized in order to decrease blind spots in the surveillance area by the proposed model in Fig.3. As shown in Fig.3, without being dependent on random placement of cameras, the surveillance system organized the arrangement pattern of each camera's observation areas to cover the whole surveillance area approximately.

Additionally, we show the average of cover rate and redundant rate as calculated from 200 installation samples for a surveillance system which consists of from 6 to 18 cameras in Fig.4. The position of each camera was determined using the uniform random number. We can see from Fig.4 that the proposed system has the adaptability to organize the arrangement pattern of observation areas to decrease the blind spots in the surveillance area without increasing the redundant area observed by plural cameras without being dependent on the placement of cameras.

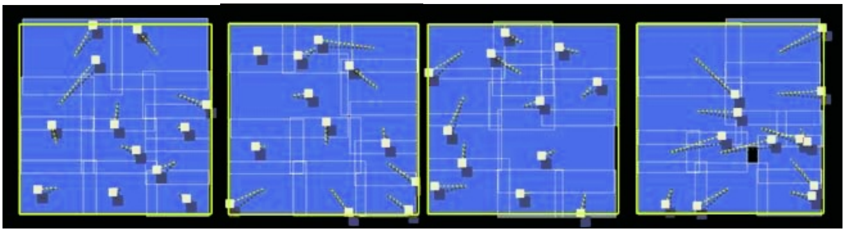


Fig. 3. Arrangement pattern of areas observed by 12 cameras installed in random position of ceiling (xy plane where z = 2.5 [m])

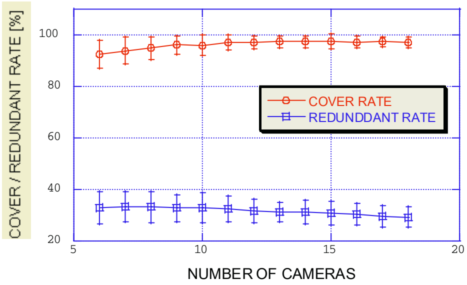


Fig. 4. Graph shows the average of cover rate and redundant rate of the surveillance system. Error bars expresses standard deviation

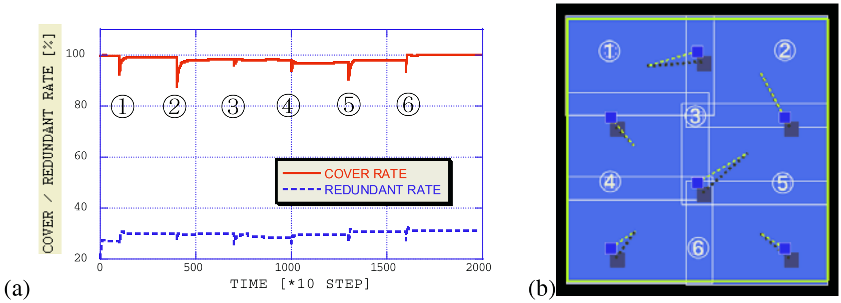


Fig. 5. (a)Graph shows the time series of the cover rate and the redundant rate of surveillance system, when some cameras are removed form the system in order. (b)Arrangement pattern of observation areas after 6 cameras are removed. The number expresses the locations where cameras are removed

Simulation 2

We show the time series of the cover rate and redundant rate of the surveillance system in Fig.5(a), when six cameras are removed one by one as shown Fig.5(b). In Fig.5(a), the time when cover rate falls greatly expresses the time when one camera is removed. Though the cover rate falls temporarily when a camera is removed, it recovers immediately since observation areas are rearranged to decrease blind spots. We can see from Fig.5 that the surveillance system has the ability to reorganize the arrangement pattern of observation area to decrease blind spots in the surveillance area, without increasing the redundant area observed by plural cameras.

5 Conclusion

In this paper, we proposed a new cooperative control model for surveillance systems in order to maintain observation of the whole surveillance area by plural cameras. Each camera adjusts their observation area to decrease blind spots around them by the control model based on reaction-diffusion equation. We have shown two simulation results which show the surveillance system has adaptability for the placement of cameras and removing of some cameras in a sense that plural cameras autonomously cooperate in order to decrease blind spots in the whole surveillance area. Although in most cases the proposed system could cover more than 90% of whole surveillance area, there are some installation samples whose cover rate is extremely low. In these cases, we consider that the proposed system is caught in the local minimum of $W(\theta_{\text{Pan}}, \theta_{\text{Tilt}}, \zeta)$ that doesn't satisfy to observe the whole surveillance area. Our future work includes improvement in avoiding such local minima which is cased by deadlock between subsystems. Moreover we will verify the practicality of our proposed system by implementing an actual system.

References

1. A. M. Turing, ”The Chemical Basis of Morphogenesis,” Philosophical Transaction of Royal Society London, vol.237B, pp.37-72, 1952.
2. Kauffman,S.A., in Pattern Formation, (eds. Malacinsky, G. M. & Bryant, S.) , pp.73-102, (Macmillan, New York), 1984.
3. Meinhardt, H., “Models of biological pattern formation,” Academic, London, 1982.
4. Kondo S: “A reaction-diffusion wave on the skin of Pomacanthus, a marine angelfish” Nature 376 765-768.
5. Hideo Yuasa, *et al.* “Self-organizing System Theory by Use of Reaction-Diffusion Equation on a Graph with Boundary,” Proc. of IEEE SMC ’99, vol.1, pp.211-216, 1999
6. Hideo YUASA, *et. al.*, Internal Observation Systems and a Theory of Reaction-Diffusion Equation on a Graph, System, Man, and Cybernetics, 1998, 1998.IEEE International Conference on, Volume:4, 11-14 oct. 1998 Pages:3669-3673 vol.4
7. Masao SUGI, *at. al.*, Autonomous Decentralized Control of Traffic Signals with Closed-Loop Constraints on Offsets, SICE Annual Conference in Fukui, August 4-6, 2003
8. Thomchana Trevai, *et. al,* Cooperative Exploration of Mobile Robots Using Reaction-Diffusion Equation On Graph. Preceedings of the 2003 IEEE International Conference on Robotics & Automation Taipei, Taiwan , September pp.14-19, 2003.
9. Norimichi Ukita and Takashi Matsuyama, ``Real-time Cooperative Multi-target Tracking by Communicating Active Vision Agents,'' Computer Vision and Image Understanding, Vol.97, No.2, pp.137-179, 2005.

Comparison of the Effectiveness of Decimation and Automatically Defined Functions

D.T. Nanduri and Vic Ciesielski

School of Computer Science and Information Technology
RMIT University, GPO Box 2467V, Melbourne, Australia
vc@cs.rmit.edu.au

Abstract. Decimation and automatically defined functions are intended to improve the fitness of the generated programs and to increase the rate of convergence to the solution. Each method has an associated computational cost, the cost for automatically defined functions being considerably higher than for decimation. This paper compares the performance improvements in genetic programming provided by automatically defined functions with that of decimation on four common benchmark problems – the Santa Fe ant, the lawnmower, even 3-bit parity and a symbolic regression problem. The results indicate that decimation provides improvement in performance that justifies the additional computation but the added computational effort required for automatically defined functions is not justified by any performance improvements.

1 Introduction

Genetic programming is a computationally intensive process. Evolving a large population of individuals through generations, using crossover and mutation, and evaluating them for each generation is a very time consuming and processor intensive task. There is a major research interest in techniques that will deliver better programs with lower computational effort. Methods such as decimation and automatically defined functions can be used for reducing the amount of processing required in genetic programming for finding the solution, usually by trying to reduce the number of evaluations. These methods themselves require some additional processing, and this processing can be justified if there is an overall improvement in performance.

Decimation refers to a genetic operation in which a large initial population is first constructed, the individuals evaluated and all but a small number of individuals deleted. The remainder of the evolutionary run is carried out with this smaller population [7]. When decimation is carried out, the additional computational effort is usually not very high, since decimation only involves creating some additional individuals, evaluating all of them and then getting rid of the bad ones. It is a one-time process and the rest of the genetic programming run remains unchanged.

While it is generally believed that decimation is a beneficial operation there are very few reports in the literature on its effectiveness. Decimation is used in [3], but its effect on the evolutionary run is not given. In the area of genetic algorithms there have been a number of studies, for example[6], which attempt to adjust the size of the population dynamically during the course of the run. This can be considered as a form of dynamic decimation. In [4] a kind of decimation over populations is described. A

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 540–546, 2005.

number of populations are evolved concurrently and the worst performing populations are pruned.

Automatically defined functions (ADFs) refers to a process of attempting to generate re-useable subroutines in the evolved programs [8]. When automatically defined functions are used, each program in the population conforms to a constrained syntactic structure. Much additional processing is required for executing the programs and creating and maintaining the structure during a run of genetic programming. An individual is likely to be more expensive to evaluate because the subroutines cound be invoked several times. The main genetic operators, crossover and mutation, need to be altered so that the structure is preserved, in order to maintain the syntactic validity of all offspring. This approach was first presented in [8] and further investigated in a number of other studies, for example [2,11].

Automatically defined functions are used in the hope that some good aspects of the evolved genetic programs are saved as subprograms and may be protected from being lost due to mutation or crossover. This could help improve the rate at which a run of genetic programming will converge to a solution, since these good aspects of the evolved programs are not easily lost and may be reused.

The general aim of this paper is to compare the performance improvements of genetic programming provided by decimation with that of automatically defined functions. In particular we address the following research questions:

1. Does decimation provide any improvement in the performance of genetic programming?
2. Do automatically defined functions provide any improvement in the performance of genetic programming?
3. Is the additional computation required by the decimation and automatically defined functions justified by any improvement in performance?

We have used test problems ranging from simple to complex and problems with varying amounts of regularity. We expect automatically defined functions to perform better on problems with more regularity, the ant and the lawnmower, since they can exploit this to find the solution faster. We do not expect decimation to be affected by the type of problem.

The test problems used for this paper are Santa Fe Ant Problem [9], the lawnmower problem [8], even-3-parity problem [8] and symbolic regression [8]. The function used for symbolic regression is $3x^2/(x+2) +4x-127$.

2 Experiments and Results

Four configurations of genetic programming were explored for each problem: (1) Normal, that is, no decimation or automatically defined functions (2) Decimation only (3) ADFs with a limit of one automatically defined function (4) both decimation and ADFs. Decimation was carried out after 1 generation. The parameter values used are shown in table 1. Fifty runs were carried out for each problem.

The average best fitness vs number of evaluations over the 50 runs for each method on the Santa Fe ant problem is shown in Figure 1. Note that the curves involving decimation start after the initial generation has been evaluated and the decimation operation completed.

Table 1. Parameter values for runs. Number of runs: 50

	Pop size before decimation	Population size	Max Gen	Xover	Mut	Elitism
Santa Fe Ant	5,000	1,000	200	70%	25%	5%
Lawn Mower	5,000	500	100	70%	25%	5%
3-Parity	2,000	200	100	70%	25%	5%
Symbolic Regression	800	100	500	70%	25%	5%

In Figure 1, it can be seen that, for the Santa Fe ant problem, improvement in fitness is fastest in the case of decimation and the normal case followed by the runs where automatically defined functions are used. When decimation was used with automatically defined functions the improvement in fitness is the slowest. T-tests performed at 200,000, 175,000 and 150,000 evaluations show decimation to be superior to automatically defined functions at a confidence level of 0.05. This pattern, where decimation is the best after some initial period, is repeated for the 3-parity problem and the symbolic regression problem.

The parity and symbolic regression problems gave similar patterns for average best fitness vs evaluations. The complete data can be found in [10]. The lawnmower problem is the exception. As can be seen from Figure 3, improvement in fitness is fastest in the case where both decimation and automatically defined functions were used, very closely followed by just decimation. Automatically defined functions alone also provide an improvement over the normal case. The good performance of ADFs in this problem is not unexpected as there is an underlying regularity, mow a square and go to another square, which can potentially be captured by an ADF.

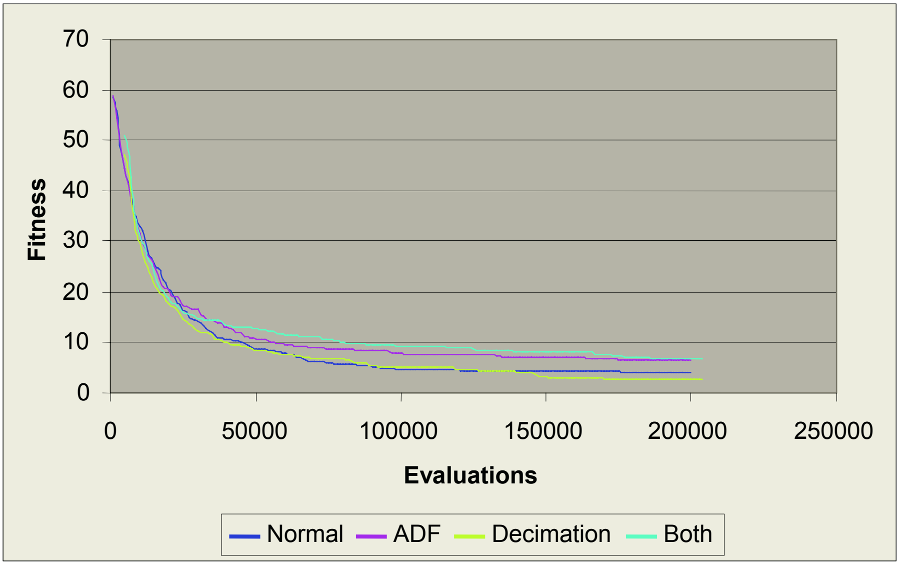


Fig. 1. Average Best Fitness for the Santa Fe Ant Problem. The parity and symbolic regression problems were similar

The number of successful runs after a given amount of CPU time for the Santa Fe and Lawnmower problems is shown in Figure 2 and 4. CPU time has been used rather than number of evaluations to make the comparison fair. Using evaluations is not fair as the evolved programs with ADFs require more expensive crossover and mutation operators and will be more costly to execute due to the subroutine structure. We have taken great care to ensure that the only differences in the programs during the runs relate to decimation and ADFs and that all runs were carried out on a dedicated machine with no other user programs running. From figure 2 it can be seen that, for example, by 110 seconds there were 31 successful runs in the normal case, 26 when decimation was used and 19 when automatically defined functions were used. The graph for the normal case, in Figure 2 is shorter than the others since there are no successful runs that took more than 130 seconds. From this graph it can be seen that, at any given point in time, there were more runs that succeeded with normal configuration than any other and that decimation is always superior to automatically defined functions. The probabilies of success, after 190 seconds generations, for normal, decimation, ADFs and both are 0.64, 0.62, 0.40 and 0.40 respectively.

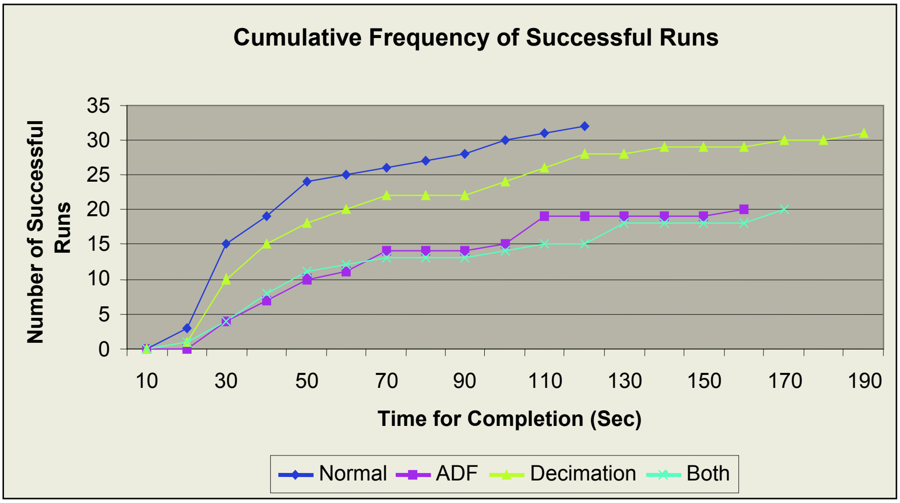


Fig. 2. Number of successful runs of the Artificial Ant problem

For the other problems, decimation is the best performer for the lawnmower and parity problems, while normal is best for the symbolic regression problem, closely followed by decimation. The complete data can be found on [10].

These results suggest that the improvement in performance given by decimation is superior to that of automatically defined functions because the runs with decimation are more likely to succeed. Also the runs with decimation converged towards a solution much faster with respect to the number of evaluations performed than did runs with ADFs.

3 Conclusions

The general goal of this paper was to compare the performance improvements to genetic programming provided by decimation with that of automatically defined functions. On all test cases decimation was better in terms of rate of improvement of fitness with evaluations and probability of getting a successful solution after expending the same amount of CPU time.

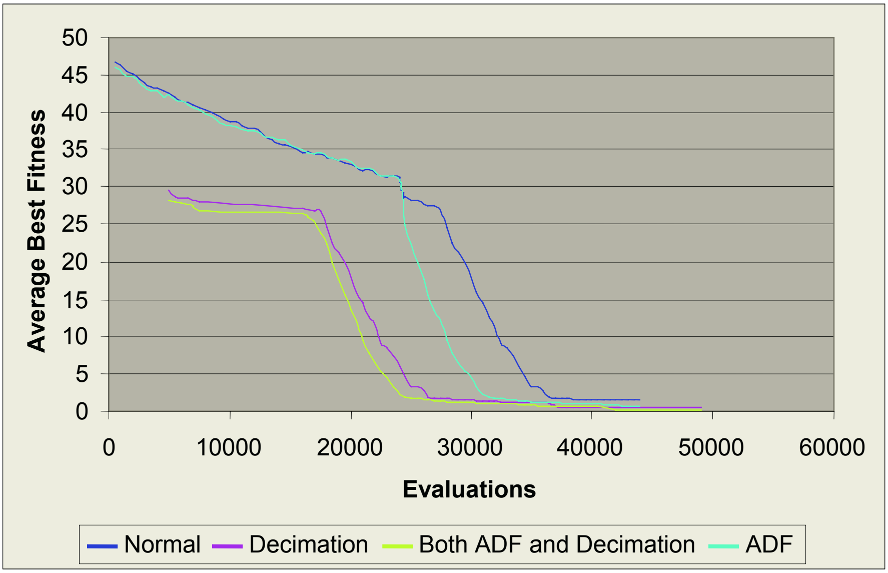


Fig. 3. Average Best Fitness for the runs of the Lawnmower Problem

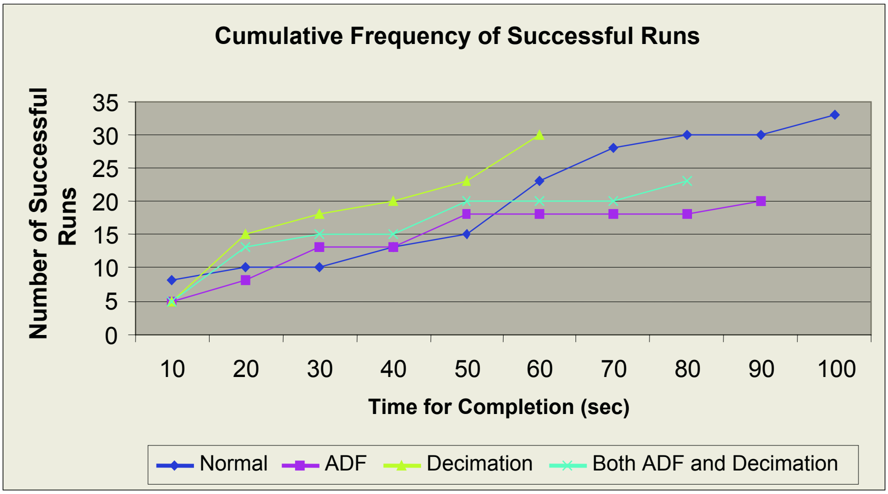


Fig. 4. Number of successful runs of the Lawnmower problem

Our first specific research question was *Does decimation provide any improvement in the performance of genetic programming?* On two of the test poblems decimation was clearly better than the normal case on both measures used. On the other two problems the performance of decimation was very close to that of the normal case. We can conclude that, on these four problems, decimation was beneficial on two problems and did not cause significant deterioration on the other two.

Our second research question was *Do Automatically Defined Functions provide any improvement in the performance of genetic programming?* On all of the test problems the performance of ADFs was worse than the normal case on both measures. The runs using automatically defined functions took much longer to succeed. The number of runs that succeeded was also lower when automatically defined functions were used. On two of the problems, using decimation together with ADFs improved performance, but not on the other two problems.

Our third research question was Is the additional computation required by decimation and automatically defined functions justified by any improvement in performance? The results from the four test problems suggest that the added computational effort needed for decimation is justified since there might be a significant improvement in the probability of getting a solution for the CPU time invested and there will not be significant deterioration.

In the case of automatically defined functions the extra computational effort is not justified. On three of the problems, the probability of getting a solution without ADFs was always higher that the probability of getting a solution with ADFs for the CPU time invested. On the other problem, this was also the case after an initial period where there was not much difference.

References

1. Aler, R. (1998) Immediate Transfer of Global Improvements to All Individuals in a Population Compared to Automatically Defined Functions for the EVEN-5, 6-PARITY Problems. *EuroGP '98: Proceedings of the First European Workshop on Genetic Programming*. Springer Verlag pages 60-70.
2. Angeline, P.J. (1996), An investigation into the sensitivity of genetic programming to the frequency of leaf selection during subtree crossover. In J. R. Koza, D. E. Goldberg, D. B. Fogel, and R. L. Riolo, editors, *Genetic Programming 1996: Proceedings of the First Annual Conference*, pages 21-29, Stanford University, CA, USA, 28-31 July 1996. MIT Press.
3. Carbajal, S., and Martinez G. (2001) Evolutive Introns: A Non-Costly Method of Using Introns in GP, *Genetic Programming and Evolvable Machines*. 2(2) Pages 111-122.
4. Ciesielski, V., and Li, X., (2003) Pyramid Search: Finding solutions for deceptive problems quickly in genetic programming. *Proceedings of the 2003 Congress on Evolutionary Computation (CEC2003).* Sarker, R. et al (Editors) IEEE Press, pages 936-943.
5. D'haeseleer, P. (1994) *Context preserving crossover in genetic programming*. In Proceedings of the 1994 IEEE World Congress on Computational Intelligence, volume 1, pages 256--261. IEEE Press.
6. Eiben, A., Marchiori, E., and Valko, V. (2004) Evolutionary Algorithms with on-the-fly population size adjustment. *Problem Solving from Nature - PPSN VIII,* Yao X. et al. (Editors), LNCS Volume 3242, Springer Verlag, pages 41-50.
7. Koza, J. (1992) *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press.

8. Koza, J.R. (1994), *Genetic Programming II: Automatic Discovery of Reusable Programs.* Cambridge, MA: The MIT Press.
9. Langdon, W., and Poli, R. (1998) Why ants are hard. *Genetic Programming 1998: Proceedings of the Third Annual Conference.* Koza, J. et al. (Editors) Morgan Kaufman, Pages 193-201.
10. Nanduri. D. T. (2005) Comparison of the Effectiveness of Decimation and Automatically Defined Functions. *Masters Thesis*, RMIT Department of Computer Science, http://www.cs.rmit.edu.au/~vc/papers/nanduri-mbc.pdf.
11. Rodrigues, E., and Pozo, A. (2002) Grammar-Guided Genetic Programming and Automatically Defined Functions. *SBIA '02: Proceedings of the 16th Brazilian Symposium on Artificial Intelligence.* Springer Verlag, pages 324-333.

Absolute Capacities for Higher Order Associative Memory of Sequential Patterns

Hiromi Miyajima and Noritaka Shigei

Kagoshima University, 1-21-40 Korimoto, Kagoshima 890-0065, Japan
miya@eee.kagoshima-u.ac.jp

Abstract. In the previous paper, we have shown the conventional capacity of higher order associative memory of sequential patterns. The definition of the capacity permits that some of the elements of the pattern are not recalled correctly. If we want to recall the patterns more correctly, then the definition is not always valid. The absolute capacity is known as the severe definition that almost all the elements of the pattern are recalled correctly. In this paper, we show the absolute capacities of higher order associative memory of sequential patterns. Further, the absolute capacities are compared with the conventional ones. Specifically, the relation between the capacity and the correlation are shown.

1 Introduction

Associative memory of sequential (time-series) patterns is an important problem in the cases where we construct associative memory system as a model of the brain and associative memory of sequential patterns is desired to apply to various type of applications[1–3]. In the previous papers, we have proposed generalized higher order associative models of sequential patterns and shown the capacities of them[4, 5]. In this case, the capacity is defined as the probability that recalling of each neuron is successful, but not as one that almost all the neurons are recalled correctly. The former is known as the conventional capacity and the latter is called the absolute capacity[1–3]. We have already shown the conventional capacities of higher order associative memory of sequential patterns[4, 5]. If we want to recall the patterns more correctly, the former is not sufficient. However, the absolute capacities of higher order associative memory of sequential patterns have been never obtained yet. In this paper, the absolute capacities for the correlation and differential correlation models are shown and compared with the conventional ones. Further, the relation between the capacity and the correlation are shown.

2 Higher Order Associative Memory of Sequential Patterns

Let us consider the conventional model consisting of N neurons mutually connected. The output for each neuron is given by

$$u_i(t) = \sum_{[l_k]} v_{i[l_k]} x_{l_1}(t) \cdots x_{l_k}(t) + b x_i(t) - \theta_i, \tag{1}$$

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 547–553, 2005.

$$\sum_{[l_k]} = \sum_{l_1} \sum_{l_2} \cdots \sum_{l_k}, \tag{2}$$

$$l_{a-1} + 1 \leq l_a \leq N - k + a, \tag{3}$$

$$x_i(t+1) = \mathrm{sgn}(u_i(t)) = \begin{cases} 1 & u_i(t) > 0 \\ -1 & u_i(t) \leq 0, \end{cases} \tag{4}$$

where $x_{l_i}(t)$ is output of the l_i-th neuron at step t, $u_i(t)$ is the internal potential of the i-th neuron at step t, $\mathrm{sgn}(u)$ is the output function, $v_{i[l_k]}$ is the weight for products of input to the i-th neuron, b is the weight for self-loop of each neuron, θ_i is the threshold of the i-th neuron, $i = 1, \cdots, N$, $t = 0, 1, \cdots$, $a = 1, \cdots, k$, $l_0 = 0$, and k is the order (dimension) of the products of the network.

In this paper, two associative memory models are considered. Let us consider sequential patterns as follows:

$$\boldsymbol{S}^1 \to \boldsymbol{S}^2 \to \cdots \to \boldsymbol{S}^P \to \boldsymbol{S}^1 \to \cdots, \tag{5}$$

where $\boldsymbol{S}^\mu = (s_1^\mu, \cdots, s_N^\mu)^T$ $(\mu = 1, \cdots, P)$ and $s_i^\mu = +1$ or -1. Each element of the patterns is selected randomly. For sequential patterns, each weight of higher order associative memory models is defined as follows:

$$v_{i[l_k]} = \begin{cases} \frac{1}{\binom{N}{k}} \sum_{\mu=1}^{P} (s_i^{\mu+1} - a)(s_{l_1}^\mu - a) \cdots (s_{l_k}^\mu - a) \\ \qquad\qquad\qquad\qquad\qquad \text{The correlation model} \\ \frac{1}{\binom{N}{k}} \sum_{\mu=1}^{P} (s_i^{\mu+1} - s_i^\mu)(s_{l_1}^\mu - s_{l_1}^{\mu-1}) \cdots (s_{l_k}^\mu - s_{l_k}^{\mu-1}) \\ \qquad\qquad\qquad\qquad\qquad \text{The differential correlation model}, \end{cases} \tag{6}$$

where a is constant.

In order to get the transition properties of the models, four assumptions for the models are made as follows:

1. Each element s_i^μ of the sequential patterns is as follows:

$$\Pr\{s_i^\mu = -1\} = p, \tag{7}$$

   where $\Pr\{\cdot\}$ is the probability of the event $\{\cdot\}$. If $p \neq 0.5$, each pattern is correlative to each other.
2. P and N are sufficiently large.
3. All states s_i^μ's with different values of i and μ are mutually independent.
4. Let $\theta_i = 0$ and $b = 0$ for the correlation model and $\theta_i = 0$ and $b = 1$ for the differential correlation model.

Let $\bar{s} = 1 - 2p$. The value $\bar{s}$ means the rate of correlation. For example, if $p = 0.5$, $p = 0.4$ and $p = 0.3$, then $\bar{s} = 0$, 0.2 and 0.4, respectively. Let $a = \bar{s}$ in the correlation model. Let the pattern ratio be defined as follows:

$$r_k = \frac{P}{\binom{N}{k}}, \tag{8}$$

Table 1. The absolute capacity for the correlation model ($k = 1$)

$N \setminus \bar{s}$	0	0.05	0.1	0.15	0.20
10	0.217	0.196	0.176	0.157	0.139
10^2	0.109	0.098	0.088	0.078	0.069
10^3	0.072	0.065	0.059	0.052	0.046
10^4	0.054	0.049	0.044	0.039	0.035
10^5	0.043	0.039	0.035	0.031	0.028
10^6	0.036	0.033	0.029	0.026	0.023

where k is dimension, P is the number of sequential patterns, and N is the number of neurons. The pattern ratio means the ratio of the number of memorized patterns per one weight. The storage capacity of the model is defined as critical pattern ratio. It means how many patterns are memorized in the model.

The distance (similarity) between two patterns $\boldsymbol{S}^{\mu+t}$ and $X(t)$ is defined as follows:

$$d_t = \frac{1}{N} \sum_{i=1}^{N} s_i^{\mu+t} x_i(t), \tag{9}$$

where $\boldsymbol{S}^{\mu+t} = (s_1^{\mu+t} \cdots, s_N^{\mu+t})$, $X(t) = (x_1(t), \cdots, x_N(t))$. Then d_0 means the distance between the pattern $X(0)$ and the memorized pattern $\boldsymbol{S}^{\mu}$ and, d_1 means the distance between the pattern $X(1)$ and the memorized pattern $\boldsymbol{S}^{\mu+1}$ and so on. In this paper, the variable t as step is neglected if there does not exist any misunderstanding.

3 Storage Capacities of the Models

3.1 The Capacities for the Correlation Model

First, let us show the internal potential u_i under the assumptions 1, 2, 3 and 4. Let $X(0)$ be an input pattern similar to the memorized one $\boldsymbol{S}^{\nu}$. Then, we can get the results by computing the internal potential of $X(1)$, which is the transition pattern of $X(0)$. The following relation holds for the internal potential u_i[4]:

$$u_i = (1 - \bar{s}^2)^k \left\{ (s_i^{\nu+1} - \bar{s}) d_0^k + (1 - \bar{s}^2)^{\frac{1-k}{2}} \sqrt{r_k} h \right\}, \tag{10}$$

where h is distributed according to the standardized normal law $N(0, 1)$.

The Absolute Stability. The storage capacity R_k in the meaning of the absolute stability is defined as the critical capacity r_k satisfying in the following equation:

$$\left(\text{Prob}\left\{x_i(1) = s_i^{\nu+1}\right\}\right)^N \geq p_e \tag{11}$$

where p_e is the positive number and nearly equals 1, such as $p_e = 0.98$. It means all the elements are recalled correctly. In order to get r_k satisfying the Eq.(11), let us compute the term $\text{Prob}\{x_i(1) = s_i^{\nu+1}\}$.

$$\begin{aligned}\mathrm{Prob}\left\{x_i(1)=s_i^{\nu+1}\right\} &= \Pr\left\{s_i^{\nu+1}=1\right\}\times\Pr\left\{u_i>0|s_i^{\nu+1}=1\right\}\\ &+\Pr\left\{s_i^{\nu+1}=-1\right\}\times\Pr\left\{u_i\le 0|s_i^{\nu+1}=-1\right\}\\ &=\frac{1+\bar{s}}{2}G\left(\frac{(1-\bar{s})}{\sqrt{(1-\bar{s}^2)^{1-k}r_k}}\right)+\frac{1-\bar{s}}{2}G\left(\frac{(1+\bar{s})}{\sqrt{(1-\bar{s}^2)^{1-k}r_k}}\right)\end{aligned} \tag{12}$$

where $G(u)=\frac{1}{2\pi}\int_{-\infty}^{u}\exp(-t^2/2)dt$. The following relation holds as $u\to\infty$:

$$G(u)\leftarrow 1-\frac{1}{2\sqrt{\pi}}\exp(-\frac{u^2}{2})/\frac{u}{\sqrt{2}} \tag{13}$$

From the Eqs.(11), (12) and (13), the following result holds:

$$\log N<\frac{(1-\bar{s})^2}{2(1-\bar{s}^2)^{1-k}r_k} \tag{14}$$

Then let R_k be the maximum number satisfying the Eq.(14).

Proposition 1.

$$R_k=\frac{(1-\bar{s})^2}{2(1-\bar{s}^2)^{1-k}\log N} \tag{15}$$

If $\bar{s}=0$, then

$$R_k=\frac{1}{2\log N} \tag{16}$$

Table 1 shows the result. Results for $k=2$ and $k=3$ are neglected as they are very similar to the case of $k=1$. It holds $P=10^6$ from the Eq.(8) for $R_2=0.033$, $N=10^4$, $\bar{s}=0.20$ and $k=2$, although the results are very low. The absolute stability is a strong criterion. If we do not require absolutely precise recalling of $s_i^{\nu+1}$, but require only recalling that is sufficiently close to the memorized $\boldsymbol{S}^{\nu+1}$, we have another definition of the capacity. It is usually used as the definition of the capacity[1–5]. We will call it the conventional capacity.

The Conventional Capacity. The storage capacity R_k in the memory of the conventional capacity is defined as the critical capacity satisfying the Eq.(17).

$$d_t\ge p_e \quad \text{for } t=1,2,\cdots, \tag{17}$$

where $p_e\approx 1$.

It is difficult to get d_t for any t, so we will predict R_k using d_1.

By using the Eq.(10), we can compute the distance d_1 between two patterns $X(1)$ and $\boldsymbol{S}^{\nu+1}$. The following relation holds for d_1[4].

Proposition 2.

$$d_1=\frac{1}{2}\sum_{\alpha\in\{-1,1\}}(1+\alpha\bar{s})\times\Phi\left(\frac{(1-\alpha\bar{s})d_0^k}{\sqrt{(1-\bar{s}^2)^{1-k}r_k}}\right), \tag{18}$$

where $\Phi(u)$ is the Gaussian function with the variable u.

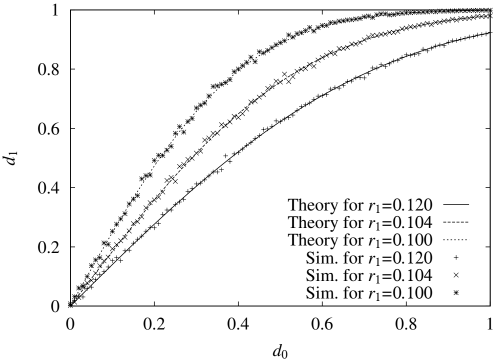


Fig. 1. Theoretical results and numerical simulations for correlative model ($k = 1$)

Therefore, Fig.1 shows numerical simulations and theoretical results for the cases of $N = 2000$, $r_1 = 0.1$, 0.104, 0.12 and $\bar{s} = 0$. The results in numerical simulations are in fairly general agreement with the theoretical ones. Let us predict the storage capacity for higher order correlation model by using the Eq. (18). First, let us assume that the case where d_1 for $d_0 = 0.98$ is greater than 0.98, is successful in recalling. Furthermore, let us define that the storage capacity is r_k for the critical case in these successful cases. Then, the storage capacities, $r_1 = 0.177$ for $k = 1$, $r_2 = 0.170$ for $k = 2$ and $r_3 = 0.163$ for $k = 3$, are obtained for $\bar{s} = 0$. The storage capacities by numerical simulations, $r_1 = 0.149$ for $k = 1$, $r_2 = 0.124$ for $k = 2$ and $r_3 = 0.129$ for $k = 3$ are obtained for $\bar{s} = 0$[4]. Table 2 is obtained from the Eq.(18).

3.2 The Capacities for the Differential Correlation Model

Likewise, let us compute the capacities for the differential correlation model. The same definitions of the capacity are used as the case of the correlation model. We have shown the following result for the internal potential u_i[5].

$$u_i = \begin{cases} (1-\bar{s}^2)^k \Big[\big\{ 2s_i^{\nu+1} - (s_i^{\nu+2} + s_i^{\nu}) \big\} d_0^k \\ \qquad + \sqrt{(1-\bar{s}^2)^{1-k} r_k 2^{k+1}}\, h \Big] + x_i(0) \quad \text{if } k \text{ is odd,} \\ (1-\bar{s}^2)^k \Big[(s_i^{\nu+2} - s_i^{\nu}) d_0^k \\ \qquad + \sqrt{(1-\bar{s}^2)^{1-k} r_k 2^{k+1}}\, h \Big] + x_i(0) \quad \text{if } k \text{ is even.} \end{cases} \tag{19}$$

The Absolute Stability. By using the same method as the Eqs.(11), (12), (13) and (14). we can get the following result:

Table 2. The conventional capacity for the correlation model applying Eq.(18)

$k \setminus \bar{s}$	0	0.05	0.1	0.15	0.20
1	0.177	0.172	0.163	0.150	0.133
2	0.170	0.165	0.155	0.140	0.123
3	0.163	0.160	0.147	0.133	0.113

Proposition 3. *Let R_k be the maximum number satisfying the Eq.(11).*

$$R_k = \frac{\{2(1-\bar{s}^2)^k - 1\}^2}{2^{k+2}(1-\bar{s}^2)^{1-k}\log N}. \tag{20}$$

If $\bar{s} = 0$, the following relation holds:

$$R_k = \frac{1}{2^{k+2}\log N}. \tag{21}$$

Table 3 shows the result. The result shows that the capacity in the absolute stability approaches to 0 gradually as N increases.

The Conventional Capacity. Let us compute the conventional capacity as the same as the correlation model. We have already shown the following results[5].

Proposition 4.

$$\begin{aligned} d_1 =& \frac{1}{8}\sum_{\alpha,\beta,\delta\in\{-1,1\}}(1+\alpha\bar{s}^2)(1+\beta\bar{s}^2)(1+\delta d_0) \\ &\times \Phi\left(\frac{\{2-\alpha(1+\beta)\}d_0^k + \bar{b}\alpha\delta}{\sqrt{(1-\bar{s}^2)^{1-k}r_k 2^{k+1}}}\right), \quad \textit{if } k \textit{ is odd,} \end{aligned} \tag{22}$$

where $\bar{b} = 1/(1-\bar{s}^2)^k$.

$$\begin{aligned} d_2 =& \frac{1}{4}\sum_{\alpha,\delta\in\{-1,1\}}(1+\alpha\bar{s}^2)(1+\delta d_0) \\ &\times \Phi\left(\frac{(1-\alpha)d_0^k + \bar{b}\alpha\delta}{\sqrt{(1-\bar{s}^2)^{1-k}r_k 2^{k+1}}}\right) \quad \textit{if } k \textit{ is even,} \end{aligned} \tag{23}$$

Here, let us predict the storage capacity for higher order correlation model by using the Eqs.(22) and (23). Then, the storage capacities, $r_1 = 0.059$ for $k = 1$, $r_2 = 0.023$ for $k = 2$ and $r_3 = 0.014$ for $k = 3$, are obtained for $\bar{s} = 0$. The relation among d_1, $\bar{s}$ and r_k has been shown in the previous paper[5].

4 Conclusions

In this paper, we have shown that the absolute capacities for higher order associative correlation and differential correlation models of sequential patterns

Table 3. The absolute capacity for the differential correlation model

(a) For $k = 1$

$N \setminus \bar{s}$	0	0.05	0.1	0.15	0.20
10	0.054	0.054	0.052	0.050	0.046
10^2	0.027	0.027	0.026	0.025	0.023
10^3	0.018	0.018	0.017	0.017	0.015
10^4	0.014	0.013	0.013	0.012	0.011
10^5	0.009	0.009	0.009	0.008	0.008
10^6	0.036	0.033	0.029	0.026	0.023

(b) For $k = 2$

$N \setminus \bar{s}$	0	0.05	0.1	0.15	0.20
10	0.027	0.027	0.025	0.022	0.019
10^2	0.014	0.013	0.012	0.011	0.009
10^3	0.009	0.009	0.008	0.007	0.006
10^4	0.007	0.007	0.006	0.006	0.005
10^5	0.005	0.005	0.005	0.004	0.004
10^6	0.005	0.004	0.004	0.004	0.003

(c) For $k = 3$

$N \setminus \bar{s}$	0	0.05	0.1	0.15	0.20
10	0.014	0.013	0.012	0.010	0.007
10^2	0.007	0.007	0.006	0.005	0.004
10^3	0.005	0.004	0.004	0.003	0.002
10^4	0.003	0.003	0.003	0.002	0.002
10^5	0.003	0.003	0.002	0.002	0.001
10^6	0.002	0.002	0.002	0.002	0.001

are $\frac{(1-\bar{s})^2}{2(1-\bar{s}^2)^{1-k}\log N}$ and $\frac{\{2(1-\bar{s}^2)^k-1\}^2}{2^{k+2}(1-\bar{s}^2)^{1-k}\log N}$, respectively. It means that the capacity approaches to 0 gradually as N increases. Further, it was clarified that two models have few affection for the low correlation.

References

1. Hertz, J., Krogh, A., Palmer, R.G., Introduction to the Theory of Neural Computation. Perseus Books Publishing (1991)
2. Amari, S.: Statistical Neurodynamics of Various Versions of Correlation Associative Memory. Proceedings of IEEE conference on Neural Networks **I** (1988) 633–640
3. Okada, M.: Notions of Associative Memory and Sparse Coding. Neural Networks **9** (1996) 1429–1458
4. Hamakawa, Y., Miyajima, H., Shigei, N., Tsuruta, T.: On Some Properties of Higher Order Correlation Associative Memory of Sequential Patterns. Journal of Signal Processing **8** (2004) 225–234
5. Miyajima, H., Shigei, N., Hamakawa, Y.: Higher Order Differential Correlation Associative Memory of Sequential Patterns. IJCNN 2004 **II** (2004) 891–896

Finding Hidden Hierarchy in Reinforcement Learning

Geoff Poulton[1], Ying Guo[1], and Wen Lu[2]

[1] Autonomous Systems, Information and Communication Technology Centre, CSIRO
PO Box 76, Epping NSW 1710, Australia
{Geoff.Poulton,Ying.Guo}@csiro.au
[2] University of NSW, Australia
wtlu159@cse.unsw.edu.au

Abstract. HEXQ is a reinforcement learning algorithm that decomposes a problem into subtasks and constructs a hierarchy using state variables. The maximum number of levels is constrained by the number of variables representing a state. In HEXQ, values learned for a subtask can be reused in different contexts if the subtasks are identical. If not, values for non-identical subtasks need to be trained separately. This paper introduces a method that tackles these two restrictions. Experimental results show that this method can save the training time dramatically.

1 Introduction

It is known that reinforcement learning (RL) methods can be extremely slow to learn when applied to problems with a large task space [1]. Dividing the problem into simpler sub-problems or using a hierarchical learning structure are ways of overcoming this. In many situations the required task of an intelligent system (e.g. a sensor network or robot) includes some repetition of sub-tasks. Such repetition is of considerable practical importance in allowing fast and efficient learning of complex tasks in very large systems. One way to benefit from such repetition is to use some form of hierarchical design with different "levels", each learning to perform a more abstract task than the level below it and providing objectives for the lower level subtask. Early examples of such methods include Feudal RL [2] and Hierarchical planning [3]. These methods require an external agency to specify the hierarchy to be used.

A more successful method is HEXQ [4], which can automatically discover the hierarchy of sub-tasks or structures in a large system. It can then identify and learn reusable sub-tasks and employ them to efficiently learn higher level skills. In HEXQ the automatic decomposition relies on particular structure in the state representation. An example is for there to be more than one variable in the state vector representing key features in the environment, where some of the variables change more slowly than others. The states in each sub-task must be represented consistently and the substructures should be identical. If these constraints are not satisfied exactly then HEXQ may not be able to automatically determine the hierarchical structure, and can only solve the flat (non-hierarchical) problem.

Unfortunately, identifying rapidly-changing variables to construct the hierarchy is not always possible in practice. We develop a novel method, Graph-Based HEXQ,

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 554–561, 2005.

which can still solve the automatic hierarchical decomposition problem under the following more difficult conditions: (1) Only one variable as the state representation; (2) the sub-tasks are similar but not identical. The method is applied to a simple but realistic experimental situation, with results which show dramatically reduced training time. Section 2 reviews HEXQ and introduces the Graph-Based HEXQ (GB-HEXQ) algorithm. Section 3 gives experimental results for a simple problem of robot navigation in a house. We conclude with some discussion on possible future research directions in Section 4.

2 Algorithm Descriptions

In this section, we will firstly explain the HEXQ algorithm and point out why it needs to be improved. We then present the GB-HEXQ algorithm in detail.

2.1 HEXQ and a Room Problem

HEXQ is a reinforcement learning algorithm that discovers hierarchical task structure automatically. It solves the problem by decomposing and discovering reusable subtasks. Parent subtasks are represented using a semi Markov decision problem (MDP) formalism. An MDP - Markov Decision Process - consists of a set of states s ; a set of actions a and a reward function R. The optimal "value" of a state is the sum of rewards received when starting in that state and following an optimal policy to a terminal state. We define the following terms for algorithm description.

- $V_m^*(s)$ is the optimal value of state s in the top-level MDP, and m_a is the sub-MDP in the level below that corresponds to an abstract action a .
- $V_{m_a}^*(s)$ is the optimal value of state s in the sub-MDP m_a .
- s_a is the exit state of the sub-MDP m_a . (Exit state will be explained in the following.)
- s' is the state the agent will be in after the exit action is taken.
- $E_m^*(s,a)$ is the optimal value state s can achieve after carrying out the abstract-action a followed by the optimal policy until the goal is achieved.
- The optimal value for s is found by choosing an abstract action a that can maximize the sum of $V_{m_a}^*(s)$ for sub-MDP m_a and its corresponding optimal state-action value $E_m^*(s,a)$.

We will illustrate the HEXQ process with a simple example where a robot must learn to navigate through a building. As shown in Fig. 1.a the building consists of three identical rooms and the goal of the robot is to get out of the building through one of the doors. The position of the robot in the building is described by two state variables, the room number (0—2) and the location-in-room (0—8).

HEXQ uses the state variables to construct the hierarchy. The maximum number of levels in the hierarchy (2) is the same as number of state variables. Firstly, the variables are ordered by their frequency of change as the robot takes exploratory random actions. In this case, the location-in-room variable changes more frequently than the room-number. The lowest level in the hierarchy is based on this variable.

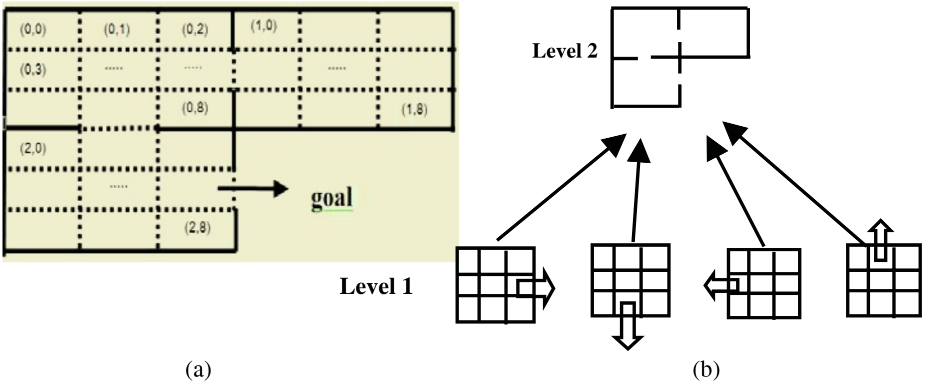


Fig. 1. (a) A building showing three rooms interconnected via doorways. Each room has 9 positions. The aim of the robot is to reach the goal. (**b**) The building is decomposed into rooms. There are 2 levels of hierarchy. In Level 1, there are 4 sub-MDPs identical except for rotation, each corresponding to one way of leaving a room

HEXQ constructs the hierarchy in a bottom-up fashion. In level 1 it uses location-in-room state transitions and reward function to see if they are invariant in all contexts defined by the room-number variable. If this is so then HEXQ can partition the state-space into regions. The boundaries of the regions are where the context changes (i.e. when room-number changes value) or where the state transitions are not invariant in different contexts. Any state-action pair that causes a transition that is not invariant is designated as an exit and that state is called the exit-state.

In the room problem, one region is just one room. Hence there are three identical regions in level one of the hierarchy, each having four exits. Fig. 1.b shows that 4 sub-MDPs, identical except for rotation, are required to find the different policies to reach each of the exits and those policies can be shared by all regions.

In constructing the second level, each region in level 1 becomes an abstract state and each policy becomes an abstract action. In this case, level 2 has 3 states and 4 actions. The second level is also the top level. There is one MDP and the exit state is simply the goal of the problem. The optimal value function for the top-level MDP can therefore be defined as [4]:

$$V_m^*(s) = \max_a [V_{m_a}^*(s) + E_m^*(s,a)] \qquad (1)$$

$$E_m^*(s,a) = R_{s_a s'}^a + V_m^*(s') \qquad (2)$$

2.3 Graph-Based HEXQ

While in regular HEXQ decomposition of a problem is done variable by variable, we now introduce GB-HEXQ which decomposes the problem by finding the "bottlenecks” from the state transition diagram. As Fig. 2 shows, in the room problem bottlenecks can be found where the state transition connects two different rooms. Each of the rooms forms a region. These regions and the policies that agents use to move between them form higher level states and actions. In the next section, we discuss the general method for decomposing the problem.

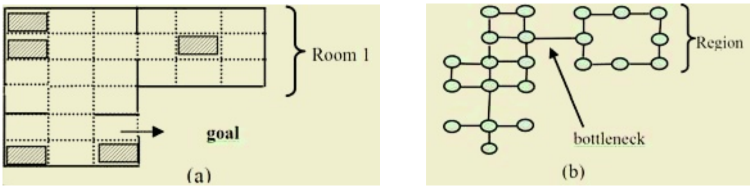


Fig. 2. (a) A modified room problem in which the shaded area is blocked by furniture and the rooms are of different size this time. **(b)** The state transition diagram of the room problem in (a)

2.3.1 Graph Partition of the State-Transition Diagrams

Fig.3 shows the state-transition diagram for two labelings of the room world. One uses a pair of variables as in HEXQ - (room number, location in room) - to represent the state, whereas the other uses a single variable – location in building. Both have exactly the same state-transition diagram.

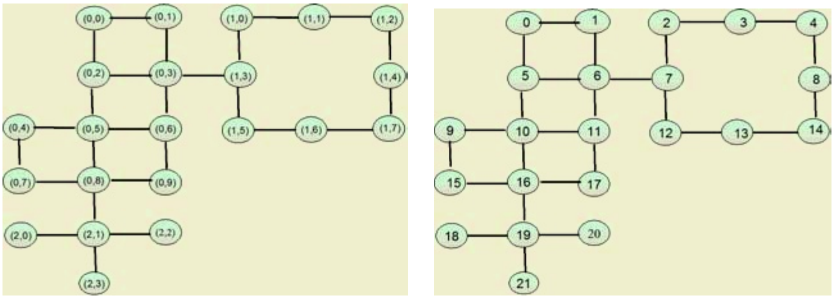


Fig. 3. Two state transition diagrams for the same room problem shown in Fig. 2 (a), each node is label with its state value

A graph partition algorithm can be used to separate the state-transition diagram into regions. In the diagram, a node represents a state, and an edge between two states shows that there exists a possible transition between them. Each edge is assumed to have unit cost. Kernighan and Lin [5] invented an algorithm that can partition the nodes of a graph with costs on its edges into subsets so as to minimize the sum of the costs on all edges that are external to the subsets[1]. The algorithm starts with two randomly chosen subsets of all nodes, then keeps swapping the elements in those two subsets until a local minimum in the total external cost is reached.

Kernighan and Lin’s algorithm can be used to split a state-transition diagram into two subsets. In order to find more than two regions, the graph partition algorithm can

[1] An external edge is an edge that connects two vertices from two different sets

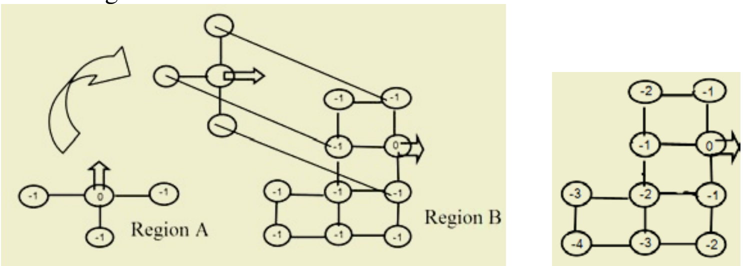
be applied again on any of the two subsets. The question is when the partition process should stop, or how do we determine whether a subset should be split again? This will be discussed in section 4. In the current implementation, the partition stops when either (a) the total external cost exceeds a threshold; or (b) either of the resulting subsets is smaller than the minimal size limit. The cost threshold and the size limit are determined by the size of the state-space. Alternatively, more sophisticated methods for graph partitioning are available, such as Hochbaum and Pathria's bottleneck partitioning [6] and Dutt's new faster Kernighan-Lin algorithm [7].

Once the state-transition diagram is partitioned into regions, the hierarchical construction of the value function is essentially the same as the HEXQ.

2.3.2 Graph Comparison – Reuse of the Values

As explained in section 2.1, in HEXQ values learned for one region can be reused in others only when their state-transitions and reward functions are identical (i.e. they are duplicates of each other under different contexts). This reuse of the value function can boost the performance in terms of training time [4].

It is obvious that reusability of the value function is also desirable when the regions are not identical. To achieve that, we first use graph comparison to find out the similarities between regions.



(a) (b)

Fig. 4. Two of the regions from the room problem. Region A can be rotated to map the nodes to Region B. **(a)** approximate values for region B from optimal values of region A; **(b)** final optimal values for region B

Graphs shown in Fig. 4 are two of the three regions of the room problem. Region A contains 4 states and Region B contains 10 states. One of exits for each region is marked by an arrow.

Conjecture: A good approximation for values in region B can be found by discovering the similar parts in region A and mapping corresponding values to B. This will lead to faster convergence for values in region B.

Fig. 4 shows a possible value mapping for region A to region B. Note that the optimal values for region B are not far from the approximation we got by mapping the values of region A to region B. Even though the regions may differ in size, shape or state-transition model, they more or less share certain features of the problem.

Observation: the value function for each sub-MDPs shares a certain pattern, which is closely related to how close a node is to the exit state and the reward it gets for each step.

To map one region to another, the exit states of both regions are always matched. Then in order to map the rest of the graph, two scalar values are used to estimate the likelihood of two nodes having the same value. The first number approximates number of steps towards the goal, and the second number is the number of adjacent nodes it has. One way to approximate steps-to-goal is to start with the exit state and expand to the nodes adjacent to it and so on until all nodes are exhausted. This is how two regions are compared.

Given more than two regions, when choosing a mapping, a scoring system is needed to find out which one will produce the best performance boost. A perfect match is when node A and node B have the same steps-to-goal and the same number of adjacent nodes. When the no perfect match can be found for a node, then it will be mapped to a node with closest values in step-to-goal and adjacent nodes. The score is calculated by counting how many perfect matches can be found.

3 Experimental Results

A house domain is used to illustrate how GB-HEXQ algorithm works. The layout of a typical residential house is shown in Fig. 5.a. There is a single robot whose task is to locate the rubbish bin.

The states are labeled randomly from 0 to 245. The robot has four primitive actions to move one cell North, South, East or West. The problem terminates when the robot reaches state 39--where the rubbish bin is located, or when it reaches state 234—the swimming pool. The actions are all deterministic, and every step that leads to a non-termination state has reward -1, when state 39 is reached, the robot receives a reward of +10 and when state 234 is reached it receives a reward of -10 (Note, the swimming pool is not a good place for a robot to go).

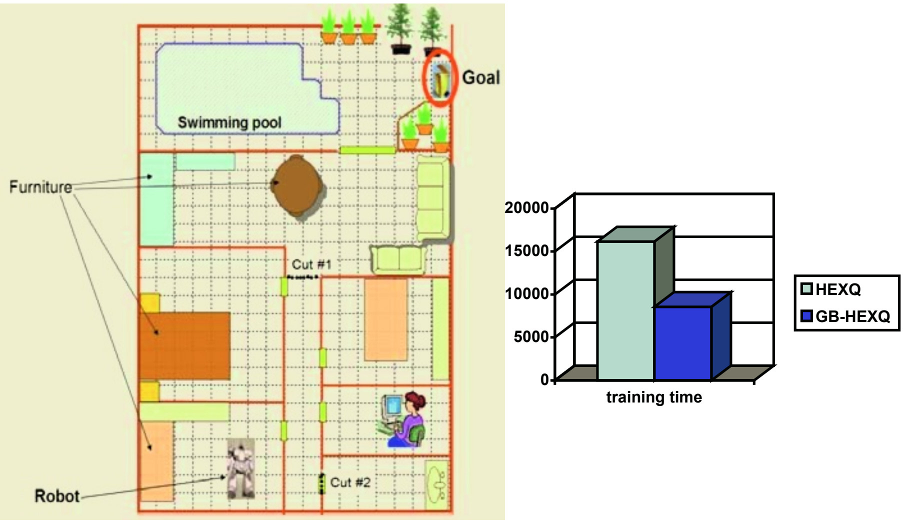
GB-HEXQ tackled this problem by first decomposing the problem into smaller pieces using the method described in section 2.3.1. For simplicity only two cuts were allowed, and the resultant division is shown in Fig. 5.a by dotted lines. Clearly other two-cut solutions are possible.

Next all the terminal and other exit states are identified. Terminal states are the goal and the swimming pool, whilst other exits are those states connected to a state belonging to a different region. Each exit state corresponds to one sub-MDP. In this particular case, because both terminal states are in the same region a total of 4 sub-MDPs must be solved in level 1. Once one of the sub-MDPs is learned, the graph comparison method described in section 2.3.2 is used to approximate the values for others before their learning starts.

In the top level hierarchy, each state corresponds to one region in the lower level, and the abstract actions available to a state are the policies lead towards the exit states of the corresponding region.

The experiment looked at how the reinforcement agents solve the house problem. Two agents are implemented in this experiment, one uses HEXQ learning, the other uses GB-HEXQ. Note that for the traditional HEXQ, when there is only one variable in the state, its performance is exactly the same as flat-Q learning.

For each trial the agent is started in a random position in the house, and allowed to wander around the house to explore the state transition pattern. Once the agent builds up a comprehensive state-transition diagram, the HEXQ learner uses dynamic programming to solve the optimal value function, while as the GB-HEXQ learner decomposes and solve the problem hierarchically. The performance of the two methods is measured by counting the number of times the *V* value needs to be looked up before the optimal value function can be found.



(a) (b)

Fig. 5. (a) A house problem, the layout of typical house, the grey area is where the furniture is. The swimming pool and the goal as marked. **(b)** Measured results on training time

As the Fig.5.b shows, GB-HEXQ saves up to about 50% of the training time compared to flat-Q learning. This is a significant gain given that the problem was split into only three parts. It is expected that further gains would be available for further splits.

4 Future Work and Conclusion

Reinforcement Learning is one of the major techniques for training robots and other autonomous systems to perform a task. HEXQ is a version which automatically discovers hierarchies in a task, leading to a significant increase in learning efficiency. In cases where the hierarchies are hidden, incomplete or both, HEXQ may not be able to automatically find out the hierarchical structure, and can only solve the flat problem. In this paper we have presented an extension of HEXQ, Graph-Based HEXQ, which can still solve the automatic hierarchical decomposition problem under the following more difficult conditions: (1) only one variable as the state representation; and (2) similar but not identical sub-tasks. GB-HEXQ has shown very encouraging results which, if borne out by further experiments, will greatly increase the range of applicable problems for reinforcement learning.

References

1. Sutton, R. S. and Barto, A. G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA, USA. (1998).
2. Dayan, P. and Hinton, G.E.: Feudal reinforcement learning. In Advances in Neural Information Processing Systems 5, S. J. Hanson, et.al. editors, pages 271-278, Morgan Kaufmann, San Mateo, CA, USA. (1993)
3. Singh, S. P.: Reinforcement learning with a hierarchy of abstract models. In Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA, USA. (1992)
4. Hengst, B.: Discovering Hierarchy in Reinforcement Learning with HEXQ. In Maching Learning: Proceedings of the Nineteenth International Conference on Machine Learning, 2002. (2003)
5. Kernighan, B.W., Lin, C.: An Efficient Heuristic Procedure for Partitioning Graphs, Bell Systems Technology J., vol. 49, no. 2, pp. 292-370. (1970)
6. Hochbaum, D.S., Pathria, A.: The bottleneck graph partition problem. Networks 28(4): 221-225 (1996)
7. Dutt, S.: New Faster Kernighan-Lin-Type Graph-Partitioning Algorithms, ICCAD '93: Proceedings of the 1993 IEEE/ACM international conference on Computer-aided design, Santa Clara, California, United States. (1993)

On-Line Reinforcement Learning Using Cascade Constructive Neural Networks

Peter Vamplew and Robert Ollington

School of Computing, University of Tasmania, Private Bag 100, Hobart
Tasmania 7001, Australia
{Peter.Vamplew,Robert.Ollington}@utas.edu.au

Abstract. In order to scale to problems with large or continuous state-spaces, reinforcement learning algorithms need to use function approximation. Neural networks are one commonly used approach, with most work so far using fixed-architecture networks. Previous supervised learning research has shown that constructive networks which grow their architecture during training outperform fixed-architecture networks. This paper extends the sarsa algorithm to use a cascade constructive network, and shows it outperforms a fixed-architecture network on two benchmark tasks.

1 Introduction

Reinforcement learning addresses the problem of an agent interacting with an environment. At each step the agent observes the current state and selects an action. The action is executed and the agent receives a scalar reward. The agent has to learn a mapping from state-to-action to maximise the long-term reward. One way to do this is to learn the expected return, either per state or per state-action pair. Many algorithms for learning these values are based on the use of temporal differences (TD) [1] where the value of the current state at each step is used to update the estimated value of previous states. For problems with small state-spaces the values can be stored in a table, but as the dimensionality or resolution of the state increases, the storage requirements become impractical. In addition learning slows, as tabular algorithms can only learn about states and actions which the agent has experienced. For these tasks function approximation must be used to estimate the values. This can usually be achieved using far fewer parameters than the number of states thereby reducing storage. In addition function approximators can generalise from states which have been experienced to similar states that are yet to be visited, hence increasing the rate of learning.

2 Reinforcement Learning Using Neural Networks

One commonly used form of function approximation is the neural network, with algorithms for both off-line [2] and on-line [3] reinforcement learning having been developed. Neural networks have been applied successfully to a range of problems, such as elevator control [4] and backgammon [5]. However this success has failed to be repli-

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 562–568, 2005.

cated on other, seemingly similar tasks. Research so far has focused on fixed-architecture multi-layer perceptrons. In this type of network, neurons are divided into layers, with connections from each layer to the next layer. The number of layers, the number of neurons per layer and the connections are all fixed prior to training. The performance of the network is dependent on the architecture and the choice of an appropriate architecture is heuristic, usually involving some form of trial-and-error.

2.1 Cascade Networks

An alternative to a fixed architecture is a constructive network, which starts from a minimal architecture, and adds neurons as needed during training. In this way the training algorithm finds an architecture suitable for the task being learnt. Many constructive algorithms have been proposed, but one of the most widely adopted has been Cascade-Correlation or Cascor [6]. Cascor has been shown to equal or outperform fixed-architecture networks on a wide range of supervised learning tasks [6], [7].

A Cascor network starts with each input connected to every output neuron, with no hidden neurons. This network is trained to minimise the mean-squared-error on a set of training examples using the Quickprop algorithm. The MSE is monitored, and if it fails to fall sufficiently over recent training epochs (as determined by a patience threshold), the decision is made to add a new hidden neuron.

A pool of candidate neurons are created, with each receiving input from all input neurons, and from any existing hidden neurons. Each candidate is trained to maximise the correlation between its activation and the residual error on the training set. Once trained, the candidate with the highest correlation is added to the network. Its weights are frozen, and it is connected to the output neurons with new random weights. Training of the output weights is now resumed. This process of alternating training of output weights and of candidates continues until a suitable solution is found or a predefined maximum number of nodes have been added.

2.2 Reinforcement Learning and Cascade Networks

[8] argues that function approximators with a bounded memory (such as a fixed-architecture network) are inherently inferior to those with an unbounded memory (such as a Cascor network) when used for reinforcement learning. Despite this justification, the only work reported using a cascade network for reinforcement learning is that of Rivest and Precup ([9], [10]), which implements an off-line learning algorithm with two alternating stages. In the first stage the agent selects and executes actions, and stores the input state and the target value generated via TD in a cache. Once the cache is full, the Cascor network is trained on the cached examples, using the standard Cascor algorithm including the use of Quickprop (Quickprop can not be used for on-line training as it is a batch training algorithm). Once the network has been trained, the cache is cleared and the algorithm returns to the cache-filling phase.

Results have been reported for tic-tac-toe, car-rental and backgammon tasks. The results for the first two tasks are promising, but the system did not perform well on the more complex backgammon task. In addition, the algorithm may not be suitable in real-time tasks due to the time requirements of the training phase.

3 Cascade-Sarsa

The algorithm proposed in this paper (cascade-sarsa) differs from that of Rivest and Precup in three main ways. First it is on-line - the network is trained after each interaction with the environment. On-line training is often preferable to off-line training, as it allows immediate learning, which is important in non-episodic tasks, or when the episode length can be extremely long (for example, where infinite loops may occur).

Second cascade-sarsa incorporates the use of eligibility traces [1] which have been shown to result in faster convergence in multi-layer perceptron training [2].

Finally it is based on a variant cascade-style network rather than Cascor. Cascor, whilst effective for classification, has some difficulties with regression due to the correlation term driving the hidden unit activations to extreme values, thereby making it hard for the network to produce a smoothly varying output. Several variants have been proposed to address this issue, using similar architectures but alternative training algorithms [11], [12], [13]. Learning state-action values is a regression task, and therefore we have used one of these variants - Cascade2, developed by Fahlman as reported in [11]. Cascade2 differs from Cascor by training candidates using the residual error as a target output. To facilitate this process, output weights are trained for each candidate in addition to its input weights, and these are also transplanted to the main network when a new hidden node is added.

3.1 Training Output Weights

As in [2] and [3] we use a separate single-output network for each action, rather than a network with an output per action. Each network starts with no hidden units, with input units connected directly to the output unit, which uses a linear activation function. The weights on these connections are trained using sarsa [3], as outlined below:

```
for each learning episode
   clear all eligibility traces
   while (! end of the episode)
      observe the current state of the environment
      calculate the output of each network
      select an action a based on networks' outputs
      $Q_t$ = output of network a
      if this action is not the first in the episode
         $\delta_{TD} = r + \gamma Q_t - Q_{t-1}$
         update weights for each network to minimise $\delta_{TD}$
         recalculate network activations
      $Q_{t-1}$ = output of network a
      update eligibility traces for all networks
      execute action a and observe the reward r
```

3.2 Adding Hidden Neurons

In parallel with the training of the output neuron's weights, candidate neurons are trained for each network. Eligibility trace and weight updates are performed for these candidates as if they were connected to the network, but they do not contribute to the activation of the output neuron. The input and output weights for the candidates are trained to reduce the residual error, by minimising the following term, where w_c is the output weight for the candidate neuron, and o_c is the current activation of the candidate neuron:

$$\delta_C=\delta_{TD} - w_c o_c \tag{1}$$

In Cascor, the output error is accumulated during training, and periodically tested against the patience threshold to decide whether to add a new node. In sarsa the same error (δ_{TD}) is used for all networks so using this term in patience tests would result in the same topology for all networks. The function to be learnt may vary in complexity between the different actions, and so each network should be able to grow its own topology. Hence cascade-sarsa's patience testing is based on the following term which weights the error by the eligibility traces of the output neuron's weights, where n is the number of weights for this output, and e_w is the eligibility trace for weight w:

$$\sum_{w=1}^{n}\left(\delta_{TD}e_w\right)^2 \tag{2}$$

Similarly, the term in equation 3 is accumulated for each candidate (where m is the number of input weights for this candidate), and when a new hidden node is added the candidate with the lowest accumulated value is selected:

$$\sum_{w=1}^{m}\left(\delta_c e_w\right)^2 \tag{3}$$

Once a new hidden neuron is added, its input weights are frozen and so there is no need to maintain eligibility traces or perform weight updates for these weights.

4 Experimental Method

Cascade-sarsa was compared against a fixed-structure network on two benchmark problems from the RL literature - Puddleworld and Mountain-Car [14], [15]. These problems were chosen as [14] reports that a fixed-structure network was unable to learn a suitable policy on either of these tasks.

As shown in Figure 1(a), Puddleworld is a two-dimensional environment. The agent starts each episode at a random, non-goal state and has to find its way to the goal. The agent receives its current coordinates as input, and at each step selects between four actions (left, right, up or down) which move it by 0.05 in the desired direction. At each step a small amount of gaussian noise (standard deviation 0.01) is also added. The agent's position is bounded by the limits of the world (0...1). On each step on which the goal is not reached, the agent receives a penalty of -1. An additional penalty is applied when the agent is within a puddle, equal to 400 multiplied by the distance to the nearest edge of the puddle.

As shown in Figure 1(b), the Mountain-Car task requires a car to escape from a 1-dimensional valley. The car's engine is less powerful than gravity, and so the car must reverse up the left-hand side of the valley to build enough potential energy to escape from the right-hand side. The inputs to the agent are the car's current position and velocity, and there are three possible actions - full throttle forward, full throttle backward, and zero throttle. As with Puddleworld a penalty of -1 is received on all steps on which the goal-state is not reached.

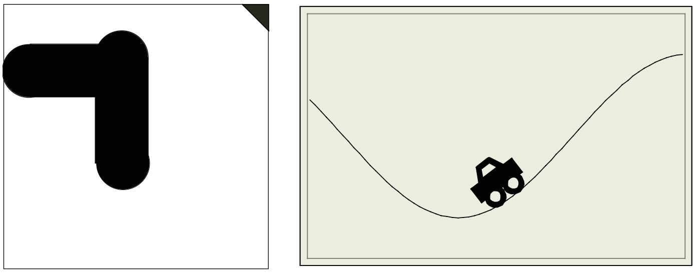


Fig. 1. (a) Puddleworld. The goal is the triangle in the top-right corner, and the puddles are capsules with radius 0.1, defined by the line segments (0.1, 0.75) to (0.45, 0.75), and (0.45, 0.4) to (0.45, 0.8). (b) Mountain-Car. The goal is to escape from the right-hand edge of the valley

For each problem, a number of trials were run for each type of network to find appropriate values for the parameters (α, λ, number of hidden nodes, and patience period length (a fixed patience threshold of 0.95 was used for the cascade-sarsa trials)). For each set of parameters 20 networks were trained, with different initial random weights. Each network was trained over 1000 episodes, using ε-greedy selection with $\varepsilon = 0.2$ and $\gamma = 1$. Each episode ended either when the goal state was reached, or after 1000 time-steps. Following training, each network's policy was assessed by running a further set of episodes using strictly greedy selection, and with learning disabled, with the starting positions sampled uniformly from the state-space.

5 Results and Discussion

Table 1 shows the results achieved by each of the network styles, for the best parameter set found for that type of network, averaged across all 20 networks trained using those parameters. In line with the findings of [14], the fixed-architecture network found these tasks difficult. In comparison cascade-sarsa was significantly more successful, particularly on the mountain-car task where it consistently learnt a near-optimal policy whereas the fixed network rarely managed to converge to a suitable policy. The difference in performance was less substantial on the Puddleworld task, although cascade-sarsa still managed to learn a superior policy, and did so using far fewer hidden neurons.

The contour plots in Figure 2 illustrate the relationship between the values of lambda and alpha and the performance of cascade-sarsa. It can be seen that cascade-sarsa is quite sensitive to these parameters, with the best results on both problems being achieved using high values of lambda.

Table 1. Results on the Puddleworld and Mountain Car tasks, for the best parameter set averaged over 20 trials. For cascade-sarsa the number of hidden nodes is the mean per network

	Fixed-architecture	Cascade-sarsa
Puddleworld		
Parameters	$\alpha = 0.005$, $\lambda = 0.5$	$\alpha = 0.001$, $\lambda = 0.9$, patience = 80
Hidden nodes	12	5.2, 4.7, 4.7, 5.2
Mean cost per test episode	298.3	257.5
Mountain-Car		
Parameters	$\alpha = 0.0005$, $\lambda = 0.6$	$\alpha = 0.001$, $\lambda = 0.9$, patience = 60
Hidden nodes	6	5.8, 6.7, 6.3
Mean cost per test episode	47.6	6.3

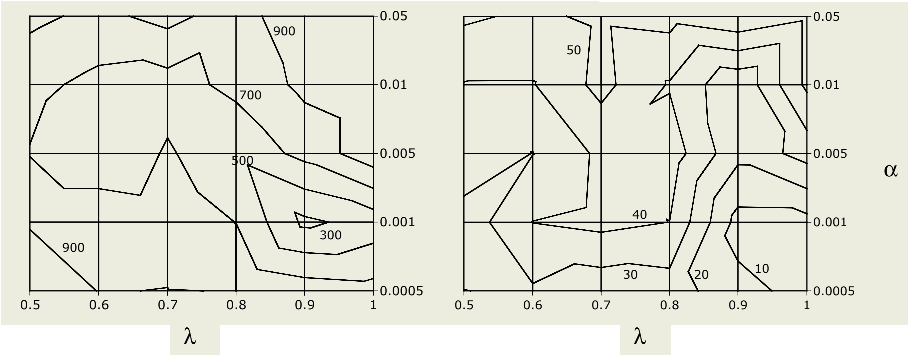


Fig. 2. Performance of cascade-sarsa (mean cost per test episode) plotted against λ and α for Puddleworld (left) and Mountain-car (right) tasks. Some values have been omitted for clarity

6 Conclusion and Future Work

The general-purpose function approximation abilities of neural networks should make them a valuable tool for use in reinforcement learning, but previous research has found the performance of fixed-architecture networks to be unreliable when trained using temporal-difference methods. This paper has presented a new algorithm based on sarsa for performing on-line reinforcement learning using a constructive network with a cascade architecture. This algorithm has been shown to outperform fixed-architecture networks on two benchmark problems. Given the experience with Cascor for supervised learning problems, it would be expected that cascade-sarsa will also scale up better than the fixed-architecture network to more difficult reinforcement learning tasks but this has yet to be experimentally confirmed.

Future work will focus on increasing the robustness and learning speed of cascade-sarsa. One area to be explored is the patience testing used in deciding when to add a new hidden node to the network. On the mountain-car task the system regularly added nodes to the ‘zero throttle’ network even though this action is redundant. We will investigate whether basing patience testing on the error of the candidate nodes rather than on the main network’s error may improve system performance. We also intend to explore the use of locally responsive neurons and competitive learning in the training of the candidates, to rapidly focus the candidates on particular portions of the residual error.

References

1. Sutton, R.S. (1988). Learning to predict by the methods of temporal differences, Machine Learning, Vol. 3, pp 9-44.
2. Lin, L. (1993), Reinforcement Learning for Robots Using Neural Networks, PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.
3. Rummery, G. and M. Niranjan (1994). On-line Q-Learning Using Connectionist Systems. Cambridge, Cambridge University Engineering Department.
4. Crites, R.H. and Barto, A.G. (1996), Improving Elevator Performance Using Reinforcement Learning, NIPS-8.
5. Tesauro, G. J. (1995), Temporal difference learning and TD-Gammon, Communications of the ACM. 38(3), pp.58-68.
6. Fahlman, S. E. and Lebiere, C. (1990). The Cascade-Correlation Learning Architecture. in Touretzky, D.S., Advances in Neural Information Processing II, Morgan Kauffman: 524-532.
7. Waugh, S.G. (1995), Extending and benchmarking Cascade-Correlation, PhD thesis, Department of Computer Science, University of Tasmania
8. Thrun, S. and Schwartz, A. (1993), Issues in Using Function Approximation for Reinforcement Learning, Proceedings of the Fourth Connectionist Models Summer School, Hillsdale, NJ, Dec 1993.
9. Rivest, F. and D. Precup (2003). Combining TD-learning with Cascade-correlation Networks. Twentieth International Conference on Machine Learning, Washington DC.
10. Bellemare, M.G., Precup, D. and Rivest, F. (2004), Reinforcement Learning Using Cascade-Correlation Neural Networks, Technical Report RL-3.04, McGill University, Canada.
11. Prechelt, L. (1997). Investigation of the CasCor Family of Learning Algorithms, in Neural Networks, 10 (5) : 885-896.
12. Adams, A. and S. Waugh (1995), Function Evaluation and the Cascade-Correlation Architecture, in Proceedings of the1995 IEEE International Conference on Neural Networks. pp. 942-946.
13. Lahnajarvi, J. J.T., Lehtokangas, M.I., Saarinen, J.P.P., (2002). Evaluation of constructive neural networks with cascaded architectures, in Neurocomputing 48: 573-607.
14. Boyan, J.A. and Moore, A.W. (1995), Generalization in reinforcement learning: Safely approximating the value function, NIPS-7.
15. Sutton R.S. (1996). Generalisation in reinforcement learning: Successful examples using sparse coarse coding. In Touretzky D.S., Mozer M.C., & Hasselmo M.E. (Eds.). Advances in Neural Information Processing Systems: Proceedings of the 1995 Conference (1038-1044). Cambridge, MA: The MIT Press.

Node Exchange for Improvement of SOM Learning

Tsutomu Miyoshi

Department of Information and Knowledge Engineering, Tottori University
Tottori-shi Koyama-cho Minami 4-101, 680-8552, Tottori, Japan
mijosxi@ike.tottori-u.ac.jp
http://mylab.ike.tottori-u.ac.jp/~mijosxi/

Abstract. Self Organizing Map (SOM) is a kind of neural networks, that learns the feature of input data thorough unsupervised and competitive neighborhood learning. In SOM learning algorithm, every connection weights in SOM feature map are initialized to random values to covers whole space of input data, however, this is also set nodes to random point of SOM feature map independently with data space. The move distance of output nodes increases and learning convergence becomes slow for this. To improve SOM learning speed, here I propose a new method, node exchange of initial SOM feature map, and a new measure of convergence, the average of the move distance of nodes. As a result of experiments, the average of the move distance of nodes comes to short that it becomes about 45%, and learning speed is improved that it becomes about 50% by this method.

1 Introduction

Many studies about Kohonen's Self Organizing Map (SOM) [1] are reported in various fields. SOM learning efficiency or learning speed must be essential to put to practical use of SOM, however, the studies about SOM learning algorithms or some improvement methods are not a major stream. We reported [3] that, order of learning data influence to SOM learning speed. Through this study, I suspected that, initial SOM feature map also influence to SOM learning efficiency or learning speed.

In this paper, to improve SOM learning speed, here I propose a new method, node exchange of initial SOM feature map, and a new measure of convergence, the average of the move distance of nodes.

Hereafter, Chapter 2 explains SOM, Chapter 3 describes about node exchange and the measure of learning, Chapter 4 describes experiments, and Chapter 5 describes a conclusion.

2 Kohonen's Self-organizing Maps

Kohonen's Self Organizing Maps (SOM) [1] is a kind of neural networks [2], that learns the feature of input data thorough unsupervised and competitive neighborhood learning. It provides a feature map that arranged similar classes in near position. SOM is mapping from a high dimensional space to usually a two dimensional space. So, it can visualize the high-dimensional information to a two dimensional feature map. Map representation makes us easy to understand the relation between data.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 569–574, 2005.

2.1 SOM Learning Algorithm

Generally, SOM has two layers, input layer and Output layer. The output layer nodes usually forms a two-dimensional grid and input layer nodes are fully connected with those at the output layer nodes. Each connection has connection weight, so every output layer nodes have pattern or vector to be learned.

In learning process, when an input pattern or input vector is presented to the input layer as learning data, the output layer nodes compete with each other for the right to be declared the winner. The winner will be the output layer node whose incoming connection weight are the closest to the input pattern in terms of Euclidean distance. The connection weight of the winner and its neighbor nodes are then adjusted i.e. moved closer in the direction of the input pattern.

As learning process progresses, learning rate and the size of the neighbor area around the winner node will be made to decrease. So, large number of output layer nodes will be adjusted strongly in the early stage of learning process, and only winner node will be adjusted weakly in the later stage. Similarly, the learning rate will decrease as learning progresses. SOM learning consists of two stages or double loop. By the inside loop, learning data is inputted in order and connection weights are adjusted. By the outside loop, learning rate and the size of the neighbor area around the winner node are made to decrease and an inside loop is repeated until learning is completed.

After learning, each node represents a group of individuals with similar features, the individual data correspond to the same node or to neighboring nodes. That is, SOM configures the output nodes into a topological representation of the original data.

2.2 Connection Weights

In initialized process, every connection weights in SOM feature map is initialized at random within the domain of each dimension. This has two remarkable meanings, one is that, initial SOM feature map covers whole space of input vectors or input data space. Another is that, this process set random point of input space into random point of SOM feature map. There is no relation between the positions of input space and feature map. Unfortunately, learning speed becomes slow is expected for this relation missing. Learning algorithm of conventional SOM shows that, neighbor area are selected based on the position of feature map in neighborhood learning, although neighbor area, that should be selected, are based on the position of input space.

As learning progress, relation between input space and feature map is self-organized, however, influence of this relation missing is serious in the early stage of learning process in particular, and the move distance of all nodes in input space increases.

3 Node Exchange and the Measure of Learning

In this chapter I examined how to exchange nodes of initial SOM feature map to make relation between the position of input space and the one of feature map without learning or without connection weight adjustment. The image of node exchange process is shown in fig.1.

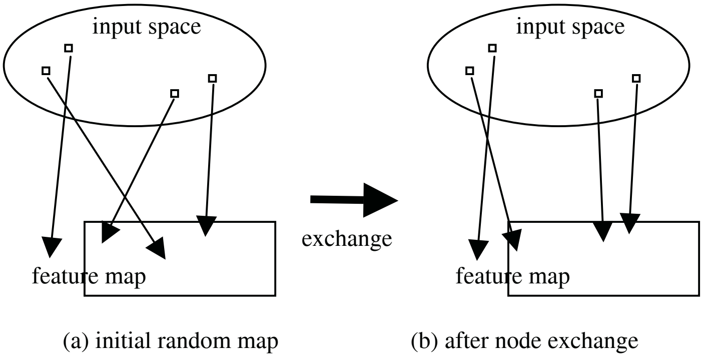


Fig. 1. Image of node exchange process

3.1 Initial Node Exchange

The concept of initial node exchange is that, the nodes, that are in the neighbor area of winner node in feature map, are exchanged to the nodes, that are in the neighbor area of winner node in input space, to make relation between position of input space and position of feature map. There are many mapping possibility from input space to feature map and desirable mapping changes depending on learning data. So, it is necessary to perform node exchange process using selected learning data.

Comparing to connection weight adjustment process and node exchange process, it can be said that node exchange process is easy to implement because these two processes are resemblance. In connection weight adjustment process, distances between learning data and all nodes are calculated in the step of selecting winner node. So it is easy to make node ordering in input space by using calculated distances. Based on the order, neighbor nodes in feature map are exchanged to neighbor nodes in input space.

The scheme of initial node exchange process is shown as following:

- 1. input vector is presented
- 2. calculate distance between input vector and all nodes
- 3. ordering nodes
- 4. set the first order node to winner node
- 5. select next order node
- 6. select a neighbor node of winner node in feature map
- 7. exchange their position in feature map
- 8. repeat 5 to 7 until all neighbor node are processed

3.2 The Measure of Learning

In conventional SOM algorithm, convergence of learning are determined by following two measures:

- 1. the number of repetition becomes larger than the threshold
- 2. the largest distance in all distances between learning data and its winner node becomes smaller than the threshold

In the 1st measure, learning convergence itself is thought as important and convergence speed is not taken into consideration. Regardless of the performance of learning algorithms, learning takes the same time or same repetition limited by threshold. Long time or large number of repetition is usually set up because it must be sufficient length for convergence of learning.

In the 2nd measure, distance between learning data and farthest winner node are used as a measure of learning. Since its attention is paid to a node with the slowest convergence, convergence of the whole feature map or whole nodes has not been measured.

In addition to these measures, I propose the average of the move distance of nodes are able to use as a new measure. It is expected that this can measure convergence of learning of the whole feature map. The average of move distance of nodes is discussed in following two cases.

Case 1: number of nodes greater than or equal to number of learning data

Case 2: number of nodes less than number of learning data

In Case 1, after enough learning, it is expected that, the move distance of nodes convergence to constant, because each learning data is assigned to unique node, and its connection weights become same value with assigned learning data.

In Case 2, while learning rate greater than 0, each node continues to move because two or more learning data are assigned to one node. It is expected, however, total move distance of nodes convergence to constant, because learning data to be assigned to the node become same members, and their influences are negated after enough learning. With same learning rate and same learning data, connection weights return to same values after performing enough repitation of adjustment.

After all, in both Case 1 and Case 2, the move distance convergence to constant, so the average of the move distance of all nodes can be used as the measure of learning convergence. It can be said that, learning becomes convergence so that this measure convergence to constant, and learning at high speed algorithm so that the move distance is small or this measure becomes smaller.

3.3 Proposed Method

Proposed Method is able to realize by inserting node exchange process between feature map initialization process and connection weight adjustment process of conventional SOM learning algorithm.

The scheme of proposed method is shown as following:

- 1. Initialize feature map at random
- 2. perform node exchange process
- 3. perform connection weights adjustment process of conventional SOM
- 4. calculate the largest distance of winner node
- 5. calculate the average of the move distance of all nodes
- 6. repeat 3 to 5 until learning is determined to convergence

4 Experiments

For the experiments, I used common parameters to compare proposed method with conventional one, 10x10 of 100 nodes two-dimensional feature map, its connection weights are initialized to random values, learning rate is gradually converged on 0.0001 from 0.01 as learning progresses, neighbor size of learning is also converged from 7x7 area to single node. Learning data are created randomly into 8 classes, the center of classes is one of the corner of 3-dimensional cube, and 20 data per class total 160 data are used.

neighbor size of node exchange, peculiar to the proposed method, is initially set to 7x7 area, and exchange process for all learning data are repeated 4 times with narrowing area such that 7x7, 5x5, 3x3, and single node.

4.1 Self-organization of Feature Map

The experiment for confirming that learning is performed satisfactory by the proposed method was conducted.

```
. 7 0 1 7 1 1 . 3 .    1 5 3 . 3 6 6 6 . .    1 1 3 3 3 6 6 6 6 .
5 2 4 . 5 5 . 7 . 1    1 1 3 3 3 6 . 6 . 1    1 1 3 3 3 6 6 6 6 .
3 4 . 5 . . . 2 7 5    1 1 . . 3 . . . . 1    1 1 3 3 3 . 6 6 6 .
0 2 1 3 . 3 6 2 1 .    1 1 . . . 1 . . . .    1 1 . . . . . . 7 .
. 7 0 5 7 4 . 0 5 .    1 . 4 . 4 . . . . 5    1 . 4 4 4 . 7 7 7 7
2 . 1 1 3 5 7 . 2 2    5 . 4 4 4 . 7 5 7 5    5 4 4 4 4 4 7 7 7 7
1 . 4 . 4 1 5 5 1 .    0 . 4 4 4 . 7 7 7 .    5 4 4 4 4 . 7 7 7 7
. . . . 6 6 1 0 6 .    5 . 5 2 . 2 7 7 7 .    5 5 . 2 2 2 . . . .
6 . 5 . 3 . . 4 . 5    5 5 5 2 2 2 0 0 . 1    5 5 . 2 2 2 0 0 0 .
2 . . 4 . 5 4 . 7 5    5 5 5 2 2 2 0 0 . 5    5 5 . 2 2 2 0 0 0 0
```

(a) initial map (b) after node exchange (c) after learning

Fig. 2. A typical feature map

A typical feature map is shown in fig.2. The number, 0 to 7, shows the node belonging to the class in it, and different number shows different class. Fig.2(a) shows that nodes are distributed at random in initial feature map. Fig.2(b) is the result of performing node exchange to initial feature map. It shows that, in spite of not performing connection weight adjustment process, the node of same class has gathered each other. Fig.2(c) shows that the self-organization of feature map is completely carried out after performing proposed method. This tendency is seen in all experiment results.

4.2 Move Distance of Nodes

The experiment for comparing how the move distance of nodes changes by the proposal method and the conventional method was conducted using feature map of the same initial connection weights and the same learning data.

The average of the move distance of nodes is shown in fig.3. In the proposed method, average movement comes to drastically short that it becomes about 45% compared with the conventional method. This means that, relation missing between the position of input space and feature map causes more than 50% of node movement in the early stage of learning process.

The distance of farthest winner node is shown in fig.4. Convergence of learning has be determined by this. In many cases, the threshold value is decided one by one depending on the characteristics of learning data. Now if threshold is set as 10, fig.3 shows that, proposed method completes learning about 50% quickly compared with the conventional method.

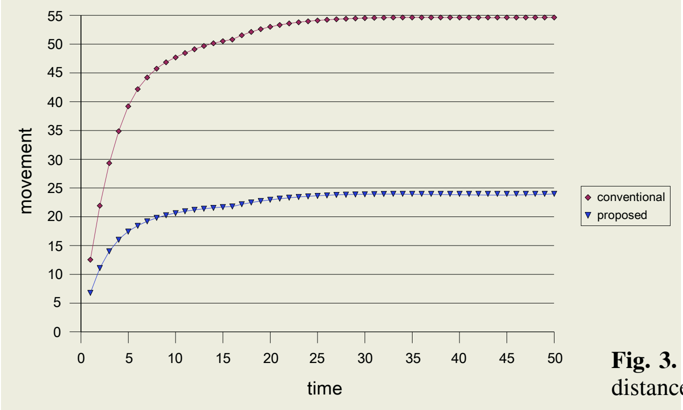


Fig. 3. The average of the move distance of nodes

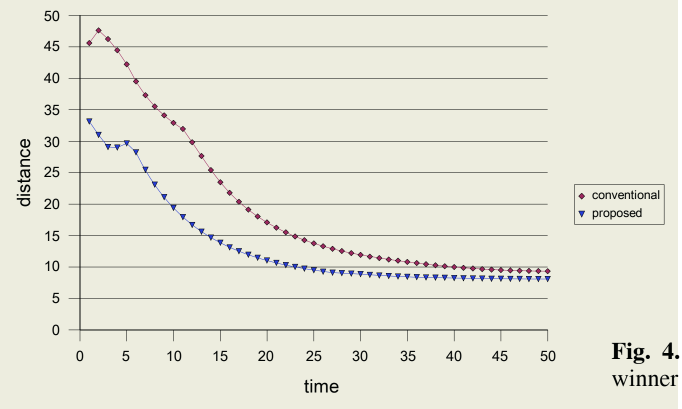


Fig. 4. The distance of farthest winner node

5 Conclusion

In this paper, I proposed a new method, node exchange of initial SOM feature map to make relation between the position of input space and the one of feature map, and a new measure of convergence, the average of the move distance of nodes to measure convergence of learning of the whole feature map.

As a result of experiments, learning speed is improved that it becomes about 50%, and the average of the move distance of nodes comes to short that it becomes about 45% by this method.

Parameter tuning of initial learning rate and initial size of neighbor, and data selection for node exchange process, etc. will be further study.

References

1. Teuvo Kohonen : "Self-Organizing Maps," Springer Verlag, ISBN 3540586008 (1995).
2. Robert Heclt-Nielsen : "Neurocomputing," Addison-Wesley Pub. Co., ISBN 0-201-09355-3 (1990).
3. Tsutomu Miyoshi, Hidenori Kawai, Hiroshi Masuyama : "Efficient SOM Learning by Data Order Adjustment," Proceedings of 2002 IEEE World Congress on Computational Intelligence (WCCI2002), USA, ISBN 0-7803-7281-6, IJCNN'02 pp.784-784 (2002).

Using Rough Set to Reduce SVM Classifier Complexity and Its Use in SARS Data Set

Feng Honghai[1,2], Liu Baoyan[3], Yin Cheng[4], Li Ping[3],
Yang Bingru[2], and Chen Yumei[5]

[1] Urban & Rural Construction School, Hebei Agricultural University
071001 Baoding, China
honghf@mail.hebau.edu.cn
[2] Information Engineering School, University of Science and Technology Beijing
100083 Beijing, China
[3] China Academy of Traditional Chinese Medicine, 100700 Beijing, China
[4] Modern Educational Center, Hebei Agricultural University
071001 Baoding, China
[5] Tian'e Chemical Fiber Company of Hebei Baoding
071000 Baoding, China

Abstract. For SVM classifier, Pre-selecting data is necessary to achieve satisfactory classification rate and reduction of complexity. According to Rough Set Theory, the examples in boundary region of a set belong to two or more classes, lying in the boundary of the classes, and according to SVM, support vectors lie in the boundary too. So we use Rough Set Theory to select the examples of boundary region of a set as the SVM classifier set, the complexity of SVM classifier would be reduced and the accuracy maintained. Experiment results of SARS data indicate that our schema is available in both the training and prediction stages.

1 Introduction

Recent results [1] indicate that the number k of SVs increases linearly with the number n of training examples. More specically,

$$k / n \to 2B_K \ . \tag{1}$$

where n is the number of training examples and B_K is the smallest classification error achievable with the SVM kernel K . When using a universal kernel such as the Radial Basis Function kernel, B_K is the Bayes risk B , i.e. the smallest classification error achievable with any decision function. Steinwart's result suggests that the critical amount of memory scales at least like $B^2 n^2$. This can be practically prohibitive for problems with either big training sets or large Bayes risk (noisy problems). Large numbers of SVs also penalize SVMs during the prediction stage, as the computation of the decision function requires a time proportional to the number of SVs. When the

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 575–580, 2005.

problem is separable, i.e. B = 0, equation (1) suggests that the number k of SVs increases less than linearly with the number n of examples. This improves the scaling laws for the SVM computational requirements.

Several techniques aim to reduce the prediction complexity of SVMs by expressing the SVM solution with a smaller kernel expansion. Since one must compute the SVM solution before applying these post-processing techniques, they are not suitable for reducing the complexity of the training stage.

Reducing the amount of training data is an obvious way to reduce the training complexity. Quantization and clustering methods might be used to achieve this goal. These methods however reduce the training data without considering the loss function of interest, and therefore sacrifice classification accuracy. Editing techniques for discarding selected training examples with the aim of achieving similar or better classification accuracy.

In this paper, we present a Rough Set based method to pre-select training examples, providing a practical means to use much larger training sets.

2 Support Vector Machine

In SVM classification [2], the samples of two classes are mapped into a feature space where they are separated by means of a maximum margin hyperplane, that is, the hyperplane that maximizes the sum of the distances between the hyperplane and its closest points in each of the two classes (the margin). The inner product in the feature space can be computed in the input space by means of a so-called kernel function. This choice is shown to affect positively the generalization performance of the classifier. When the classes are not linearly separable in the feature space, a variant of SVM, called soft-margin SVM, is used. This SVM variant penalizes misclassification errors and employs a parameter (the soft-margin constant C) to control the cost of misclassification. The support vectors lie close to the hyperplane. They therefore define both the hyper plane and the boundaries of the two classes.

3 Rough Sets

3.1 Basic Concepts

Rough Set Theory, introduced by Pawlak [3,4], involves the following: U is the universe, which cannot be empty, R is the indiscernibility relation, or equivalence relationship, A =(U, R), an ordered pair, is called an approximation space, $[x]_R$ denotes the equivalence class of R containing x, for any element x of U, elementary sets in A --the equivalence classes of R ,definable set in A --any finite union of elementary sets in A.

Therefore, for any given approximation space defined on some universe U and having an equivalence relation R imposed upon it, U is partitioned into equivalence classes called elementary sets that may be used to define other sets in A. Given that $X \subseteq U$, X can be defined in terms of the definable sets in A by the following.

lower approximation of X in A is the set $R\_(X)$={ $x \in U \mid [x]_R \subseteq X$ } ,

upper approximation of X in A is the set $R^{-}(X)$={ $x \in U \mid [x]_R \cap X \neq \Phi$ }

Another way to describe the set approximations is as follows. Give the upper and lower approximation $R^{-}(X)$ and $R\_(X)$, of X a subset of U, the R-positive region of X is $\mathrm{POS}_R(X) = R\_(X)$, the R -negative region of X is $\mathrm{NEG}_R(X) = U - R^{-}(X)$, and the boundary or R -boundary region of X is $\mathrm{BN}_R(X) = R^{-}(X) - R\_(X)$

Knowledge representation in Rough Set Theory is done via relation tables. An information system $I = \left\langle U, \Omega, V_q, f_q \right\rangle_{q \in \Omega}$ consists of (1) a finite set U of objects; (2) a finite set Ω of attributes; (3) for each $q \in \Omega$, a set V_q of attribute values, and an information function $f_q : U \to V_q$.

In the sequel we shall use I as a generic information system with $|U|$=n, and $P, Q, R \subseteq \Omega$, we also will write $f_q(x)$ to denote the value of x with respect to attribute q. Furthermore, we suppose that $d \in \Omega$ is a decision attribute that we want to predict with attributes sets $Q, R \subseteq \Omega$.

3.2 Example 1

We use the small information system given in Table 1 as an example to illustrate the various concepts developed in the sequel. An attribute "Heart Disease" (HD) shall be predicted from two variables "Smoker" (S) and "Body Mass Index" (BMI).

With each Q we associate an equivalence relation θ_Q on U by defining

$$x \equiv \theta_Q y \overset{def}{\Leftrightarrow} (\forall q \in Q) f_q(x) = f_q(y).$$

Where HD_1 and HD_2 are the equivalence classes of the HD .

Obviously the S - positive region of HD_1 or lower approximation is

$$S\_(HD_1) = S_1 = \{1,2,3,4,9\} \subseteq HD_1 = \{1,2,3,4,7,9\} \tag{2}$$

So we can hold that

$$S = no \to HD = no \tag{3}$$

Table 1. An example of an information system

No	S	BMI	HD
1	no	normal	no
2	no	obese	no
3	no	normal	no
4	no	obese	no
5	yes	normal	yes
6	yes	normal	yes
7	yes	obese	no
8	yes	obese	yes
9	no	normal	no

Table 2. The class of θ_Q and θ_d

Q	Class of θ
$\{S\}$	$\{S_1, S_2\}$={{1,2,3,4,9}, {5,6,7,8}}
$\{BMI\}$	$\{BMI_{1,} BMI_2\}$={{1,3,5,6,9}, {2,4,7,8}}
$\{HD\}$	$\{HD_1, HD_2\}$={{1,2,3,4,7,9}, {5,6,8}}

While the S - boundary region of HD_1 and HD_2 is

$$\mathrm{BN}_S(HD_1) = \mathrm{BN}_S(HD_2) = S_2 = \{5,6,7,8\} \tag{4}$$

So it can be induced that

$$S = \text{Yes} \rightarrow HD = \text{yes or } HD = \text{no} \tag{5}$$

4 Algorithm

4.1 Idea Induced from Example 1

From (2) and (3), we can conclude that if a set of any condition attribute is in the positive region of any equivalence class of the decision attributes, the objects included in the positive region belong to the correspondence class of the decision attributes determinately, while from (4) and (5) it can be induced that if a set of any condition attribute is in the boundary region of the equivalence class of any decision attributes, the objects included in the boundary region may belong to several correspondence classes of the decision attributes.

We can use the positive region to classify the objects directly. While the examples in the boundary region belong to several classes, lying in the boundary of these classes, and as we know that the support vectors lie in the boundary of two classes too, we can use SVM to classify them. In this way, the examples in the boundary region of every set are selected as the training sets singly, and the SVMs complexity is reduced.

For examples 1,2,3,4,9 in Table 2, since their class is "no", we classify them with the positive region of Rough Set Theory, while for examples 5,6,7,8, since their classes are "yes" or "no", they lie in the boundary of the two classes, we use SVM to classify them, thus the training set of SVM has been reduced to {5,6,7,8} from {1,2,3,4,5,6,7,8,9}.

4.2 Algorithm

(1) Select the condition attribute which can be separated into several equivalence classes and some subsets of them may be in positive region of any equivalence class of the decision attribute, i.e. they belong to any class determinately. While some of them may belong to several classes, namely they are in the boundary region of these classes.

(2) For the examples in positive region, use the Rough Set Theory rules to classify them.

(3) For the examples in boundary region, select them as the SVM training set, train and classify them with SVM.

4.3 SARS Data Experiment and Results

We have obtained clinical data from 524 SARS patients from Beijing. And when handling SARS data, we find the following rules (Table 4)

(1) T<=37.6°C →1 (1 means that the state of illness is slight).

(2) 37.7°C <=T<=38.6°C →1 or 2(1 or 2 means that the state of illness is slight or serious).

(3) T>=38.7°C →1, 2 or 3 (1, 2 or 3 means that the state of illness is slight, serious or critical).

Where T denotes the highest body temperature.

The above rules mean that:

The examples (T<=37.6°C) belong to the positive region of the set in which the examples' states are 1. So we can use this rule to classify the examples directly, and we need not train and classify them using SVM.

The examples (37.7°C <=T<=38.6°C) belong to the boundary region of the sets in which the examples' states are 1 or 2. So we can select these examples as a SVM training set of classes 1 and 2, accordingly the number of the training set is reduced to 103 from 492.

The examples (T>=38.7°C) belong to boundary region of the sets in which the example's state are 1, 2 or 3. So we can select these examples as training set of class 1, 2 and 3, accordingly the number of the training set is reduced to 363 from 492.

The experiments are done using the LIBSVM [5] software package. We make a random split of the data for each above pre-selecting sets into 90% training and 10% test instances. Model selection is done by 5-fold cross validation with exhaustive parameter search on the training data. The best parameter combination is then applied on the test data.

Table 3 lists the experiment results in our implementation.

Table 3. SVM experimental results of SARS data

Training set	Original Training set	T<= 37.6˚C	37.7<=T<=38.6˚C	T>=38.7˚C
Number of Examples	492	26	103	363
Class	1,2 and 3	1	1,2	1,2 and 3
Accuracy	86.7%	100%	87.9%	87.2%

In Table 3, column 2 is the original SVM data set handled by LIBSVM and selected from total of 524 examples. Column 3 is the data set classified with Rough Set Theory rules directly and the classification accuracy is up to 100%. Column 4 is the SVM training and test data set in which the examples satisfy the corresponding qualification. And so is the column 5.

5 Discussion and Future Work

In Table 3, clearly, pre-selecting training data has not only reduced the training and classification complexity but also improved the classification performance slightly.

Overall, the experimental results support the following conclusion: Rough Set based methods to pre-select examples provides a practical means to use much larger training sets. Our future work is using this algorithm to handle a much larger SARS data set.

References

1. Steinwart: Sparseness of Support Vector Machines.Some Asymptotically Sharp Bounds. In Thrun, S., Saul, L, and Sch¨olkopf, B., editors, Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA.
2. V.N. Vapnik: Statistical Learning Theory. John Wiley & Sons, (1998).
3. J.Grzymala-Busse: Managing Uncertainty in Expert Systems. Kluwer Academic Publisher, Boston, (1991).
4. Z. Pawlak. Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publisher, Norwell, MA, (1991).
5. Chang. C. & Lin, C. (2001). LIBSVM: a library for support vector machines. Software is available for download at http://www.csie.ntu.edu.tw/~cjlin/libsvm

A SVM Regression Based Approach to Filling in Missing Values

Feng Honghai[1,2], Chen Guoshun[3], Yin Cheng[4],
Yang Bingru[2], and Chen Yumei[5]

[1] Urban & Rural Construction School, Hebei Agricultural University
071001 Baoding, China
honghf@mail.hebau.edu.cn
[2] Information Engineering School, University of Science and Technology Beijing
100083 Beijing, China
[3] Ordnance Technology Institute, Shijiazhuang
050000 Shijiazhuang, China
[4] Modern Educational Center, Hebei Agricultural University
071001 Baoding, China
[5] Tian'e Chemical Fiber Company of Hebei Baoding
071000 Baoding, China

Abstract. In KDD procedure, to fill in missing data typically requires a very large investment of time and energy - often 80% to 90% of a data analysis project is spent in making the data reliable enough so that the results can be trustful. In this paper, we propose a SVM regression based algorithm for filling in missing data, i.e. set the decision attribute (output attribute) as the condition attribute (input attribute) and the condition attribute as the decision attribute, then use SVM regression to predict the condition attribute values. SARS data set experimental results show that SVM regression method has the highest precision. The method with which the value of the example that has the minimum distance to the example with missing value will be taken to fill in the missing values takes the second place, and the mean and median methods have lower precision.

1 Introduction

Because of the "garbage in, garbage out" principle, data quality problems can be very expensive - "losing" customers, "misplacing" billions of dollars worth of equipment, misallocated resources due to glitches forecasts, and so on. Solving data quality problems typically requires a very large investment of time and energy - often 80% to 90% of a data analysis project is spent in making the data reliable enough so that the results can be trustful.

Data in the real world are often plagued by missing, ambiguous values that can greatly hinder some types of analysis. There are many schemes for guessing the identity of such values, for example, using logistic regression [1], by assuming that the missing points are the same as their nearest neighbors or the same as the most abundant data type within some radius.

Complete-case analysis [2], where cases with missing values are discarded, is often conducted because its simplicity and the comparability of univariate statistics. How-

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 581–587, 2005.

ever, discarding incomplete cases may lead to a considerable loss of information and, moreover, to serious biases in estimates [2]. Means and regression imputation [2,3,4] are widely used, due to their quickness and simplicity and lack of easy-to-use software packages that implement more advanced methods, such as EM imputation [2,3,4]. For example, numbers describing the central tendency (for example mode, median or mean) have often been used in machine learning studies to treat missing values.

Recently, K-Nearest Neighbor (KNN), sample mean imputation (SMI) [7-9], multivariate regression [10-12], mixture of principal component analyzers (MPCA) and variation Bayes (VB) etc data mining methods have been introduced to carry out the imputation of missing data.

In this paper we propose a SVM regression based algorithm to fill in missing data, i.e. set the decision attributes (output or classes) as the condition attributes (input attributes) and the condition attributes as the decision attributes, so we can use SVM regression to predict the missing condition attribute values. The SARS data experiments show that our methods are available.

2 Support Vector Machine [5]

Support Vector (SV) machines comprise a new class of learning algorithms, motivated by the results of the statistical learning theory. SV regression estimation seeks to estimate functions

$$f(\mathbf{x}) = (\mathbf{w} \cdot \mathbf{x}) + b, \qquad \mathbf{w}, \mathbf{x} \in \mathbb{R}^N, \qquad b \in \mathbb{R} \tag{1}$$

based on data

$$(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_l, y_l) \in \mathbb{R}^N \times \mathbb{R}, \tag{2}$$

by minimizing the regularized risk functional

$$\|\mathbf{W}\|^2 / 2 + C \bullet R_{\text{emp}}^{\varepsilon}. \tag{3}$$

where C is a constant determining the trade-off between minimizing the training error, or empirical risk

$$R_{\text{emp}}^{\varepsilon} = \frac{1}{l} \sum_{i=1}^{l} \left| y_i - f(\mathbf{x}_i) \right|_{\varepsilon}$$

and the model complexity term $\|\mathbf{W}\|^2$. Here, we use the so-called ε -insensitive loss function

$$\left| y - f(\mathbf{x}) \right|_{\varepsilon} = \max \left\{ 0, \left| \text{y-} f(\mathbf{x}) \right| - \varepsilon \right\}$$

The main insight of the statistical learning theory is that in order to obtain a small risk, one needs to control both training error and model complexity, i.e. explain the

data with a simple model. The minimization of Eq. (3) is equivalent to the following constrained optimization problem (Vapnik, 1995):
minimize

$$\tau(\mathbf{w},\xi^{(*)}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{4}$$

subject to the following constraints

$$((\mathbf{w} \bullet \mathbf{x}_i) + b) - y_i \leq \varepsilon + \xi_i \tag{5}$$

$$y_i - ((\mathbf{w} \bullet \mathbf{x}_i) + b) \leq \varepsilon + \xi_i^* \tag{6}$$

$$\xi_i^{(*)} \geq 0, \qquad \varepsilon \geq 0 \tag{7}$$

As mentioned above, at each point $\mathbf{x}_i$ we allow an error of magnitude ε Errors above ε are captured by the slack variables ξ^* (see constraints (5) and (6)). They are penalized in the objective function via the regularization parameter C chosen a priori (Vapnik, 1995).

In the ν -SVM the size of ε is not defined a priori but is itself a variable. Its value is traded off against model complexity and slack variables via a constant $\nu \in (0,1]$
minimize

$$\tau(\mathbf{W},\xi^{(*)},\varepsilon) = \frac{1}{2}\|\mathbf{W}\|^2 + C \bullet (\nu\varepsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)) \tag{8}$$

subject to the constraints (5)–(7). Using Lagrange multipliers techniques, one can show (Vapnik, 1995) that the minimization of Eq. (4) under the constraints (5)–(7) results in a convex optimization problem with a global minimum. The same is true for the optimization problem (8) under the constraints (5)–(7). At the optimum, the regression estimate can be shown to take the form

$$f(\mathbf{x}) = \sum_{i=1}^{\ell}(\alpha_i^* - \alpha_i)(\mathbf{x}_i \bullet \mathbf{x}) + b. \tag{9}$$

In most cases, only a subset of the coefficients $(\alpha_i^* - \alpha_i)$ will be nonzero. The corresponding examples $\mathbf{x}_i$ are termed support vectors (SVs). The coefficients and the SVs, as well as the offset b ; are computed by the ν -SVM algorithm. In order to move from linear (as in Eq. (9)) to nonlinear functions the following generalization can be done: we map the input vectors $\mathbf{x}_i$ into a high-dimensional feature space Z through some nonlinear mapping, $\Phi : \mathbf{X}_i \rightarrow \mathbf{Z}_i$ chosen a priori. We then solve the optimization problem (8) in the feature space Z . In that case, the inner product of the input vectors $(\mathbf{x}_i \bullet \mathbf{x})$ in Eq. (9) is replaced by the inner product of their icons in feature space Z , $(\Phi(\mathbf{x}_i)\bullet\Phi(\mathbf{x}))$ The calculation of the inner product in a high-dimensional space is computationally very expensive. Nevertheless, under general conditions (see Vapnik, 1995 and references therein) these expensive calculations can be reduced significantly by using a suitable function k such that

$$(\Phi(\mathbf{x}_i)\bullet\Phi(\mathbf{x})) = k(\mathbf{x}_i \bullet \mathbf{x}), \tag{10}$$

leading to nonlinear regression functions of the form:

$$f(\mathbf{x}) = \sum_{i=1}^{\ell}(\alpha_i^* - \alpha_i)k(\mathbf{x}_i, \mathbf{x}) + b \tag{11}$$

The nonlinear function k is called a kernel (Vapnik, 1995). In our work we use a Gaussian kernel

$$k(\mathbf{x},\mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2/(2\sigma_{\text{kemel}}^2)) \tag{12}$$

3 Algorithm for Filling Missing Data

(1) Select the examples in which there are any not missing attribute values.

(2) Set one of condition attributes (input attribute), some of whose values are missing, as the decision attribute (output attribute), and the decision attributes as the condition attributes by contraries.

(3) Use SVM regression to predict the decision attribute values.

4 Experiment and Results

The experiments are done using the LIBSVM [6] software package on SARS data. The SARS data are obtained from the analysis of microelements Zn Cu Fe Ca Mg K Na in one's body. The category or class labels are 1 and 0, where 1 denotes that the patients are infected by SARS, and 0 not infected. Some examples of the whole data set are in Table 1 and the experiment results are given in Table 2, Table 3, Table 4.

Table 1. Some examples of whole SARS data set

Class	Zn	Cu	Fe	Ca	Mg	K	Na	Class	Zn	Cu	Fe	Ca	Mg	K	Na
1	164	22.2	35.5	2212	281	153	549	0	166	15.8	24.5	700	112	179	513
1	173	8.99	36.0	1624	216	103	257	0	185	15.7	31.5	701	125	184	427
1	202	18.6	17.7	3785	225	31.0	67.3	0	193	9.80	25.9	541	163	128	642
1	182	17.3	24.8	3073	246	50.7	109	0	159	14.2	39.7	896	99.2	239	726
1	211	24.0	17.0	3836	428	73.5	351	0	226	16.2	23.8	606	152	70.3	218
1	246	21.5	93.2	2112	354	71.7	195	0	171	9.29	9.29	307	187	45.5	257
1	164	16.1	38.0	2135	152	64.3	240	0	201	13.3	26.6	551	101	49.4	141

Table 2. Experiment results of filling in attribute Ca's values

Supposing that attribute Ca's values are missing								results of guessing Ca's values based on SVM regression							
Ca	Zn	Cu	Fe	Class	Mg	K	Na	Ca	Zn	Cu	Fe	Class	Mg	K	Na
2157	209	6.43	86.9	1	288	74.0	219.8	2158.23	209	6.43	86.9	1	288	74.0	219.8
3870	182	6.49	61.7	1	432	143	367.5	3869.39	182	6.49	61.7	1	432	143	367.5
1806	235	15.6	23.4	1	66	68.9	188	1483.64	235	15.6	23.4	1	166	68.9	188

Table 3. Experiment results of filling attribute Mg's values

Supposing that attribute Mg's values are missing								results of guessing Mg's values based on SVM regression							
Mg	Ca	Zn	Cu	Fe	Class	K	Na	Mg	Ca	Zn	Cu	Fe	Class	K	Na
288	2157	209	6.43	86.9	1	74.0	219.8	287.17	2157	209	6.43	86.9	1	74.0	219.8
432	3870	182	6.49	61.7	1	143	367.5	431.80	3870	182	6.49	61.7	1	143	367.5
166	1806	235	15.6	23.4	1	68.9	188	166.10	1806	235	15.6	23.4	1	68.9	188

In Table 1, attribute "class" is the output attribute or decision attribute, "1" denotes the patient suffers from SARS. We can use standard SVM to estimate a new example's class which it belongs to.

However, if there are some missing values in an input attribute (condition attribute), the SVM method cannot be used directly. So we set the input attribute as the output attribute or decision attribute, and set the original output attribute as one of the input attributes. Finally, use SVM regression to predict the missing values.

Table 4. Comparative results of filling in attribute Mg's and Ca's values with different methods

Real values		Method (1)		Method (2)		Method (3)		Method (4)		Method (5)	
Mg	Ca	Mg	Ca	Mg	Ca	Mg	Ca	Mg	Ca	Mg	Ca
288	2157	287.17	2158.23	113.4	2511.1	108	2220	215.5	1882.5	354	2112
432	3870	431.80	3869.39	113.4	2511.1	108	2220	202.8	1546.8	428	3836
166	1806	166.10	1483.64	113.4	2511.1	108	2220	209.1	1714.6	216	1624

In Table 2, the original output attribute is the "class". If we suppose that attribute Ca's values are missing, the attribute "class" should be set as one of the input attributes, and the attribute "Ca" be set to be the output attribute, so we can use the SVM regression method to predict the missing values of attribute Ca.

In Table 4, the left two columns are real values. Method (1) denotes the SVM regression methods proposed in this paper. In Method (2), the mean of all the values of the same class will be taken to fill in the missing values. In Method (3), the median of all the values of the same class will be taken to fill in the missing values. In Method (4), the mean of the two closest neighbor values (natural order) will be taken to fill in the missing values. In Method (5), for the example with the missing value we select the example that has the minimum distance to it, and take the value of the same attribute which the missing value belongs to to fill the missing value, i.e., value of the example that has the minimum distance to the example that contains the missing value will be taken to fill in the missing value.

In Table 4, obviously, the SVM regression method has the highest precision, Method (5) takes second place in precision, and the other methods have lower precision.

5 Discuss and Future Works

(1) The experimental results indicate that the SVM regression based algorithm for filling in missing data is available.

(2) Since the support vectors influence greatly the results of regression, the training data set had best be selected to be complete, i.e., we should select enough complete examples where there are not missing data as the training data set. If there are not enough complete examples in the training data set the regression accuracy will be influenced.

(3) The regression methods give a comprehensive and average guess for the missing data, the data, which have been filled in, reflect or embody the holistic information hidden in the whole data set, and the local information may be ignored or be submerged. This is in contrast to methods such as by assuming that the missing points are the same as their nearest neighbors where the local information is taken into account, and the holistic information ignored, resulting in bigger errors.

(4) Our future works will be the followings: (1) comparative research on different algorithms for filling in missing data such as EM, ANN etc. (2) implement the experiment on large data set.

References

1. T.M. Thomas, K.R. Plymat, J. Blannin, T.W. Meade: Prevalence of Urinary Incontinence, Br. Med. J. 281 (1980) 1243-1245.
2. R.J.A. Little, D.B. Rubin: Statistical Analysis with Missing Data, Wiley, New York, (1987).
3. J.L Schafer: Analysis of Incomplete Multivariate Data, Chapman & Hall, London, (1997).
4. M.A. Hill: SPSS Missing Value Analysis 7.5, SPSS Inc., Chicago, (1997).
5. Vapnik V N: The Nature of Statistical Learning Theory. NY: Springer-Verlag, (1995).
6. Chang. C. & Lin, C: (2001). LIBSVM: a library for support vector machines. Software is available for download at http://www.csie.ntu.edu.tw/~cjlin/libsvm
7. Zhao Guanghui, Song Huazhu, Xia Hongxia, Zhong Luo: Comparison of Missing Data Estimation Methods in Satellite Information for Scientific Exploration. DCABES (2004) 278-280
8. M. H. Cartwright, M. J. Shepperd, and Q. Song: Dealing with Missing Software Project Data. 9th International Software Metrics Symposium. (2003) 154-165
9. Eduardo R. Hruschka, Estevam R. Hruschka and Nelson F. F. Ebecken: Evaluating a Nearest-Neighbor Method to Substitute Continuous Missing Values. Lecture notes in computer science (2003) 723-734
10. Th. Liehr: data Preparation in Large Real-World data Mining Projects: Methods for Imputing Missing Values. Exploratory data analysis in empirical research (2003) 248-256

11. Jau-Ji Shen; Ming-Tsung Chen: A Recycle Technique of Association Rule for Missing Value Completion. 17th International Conference on Advanced Information Networking and Applications. (2003) 526-529
12. Mehtap KANDARA, Osman KANDARA: Association Rules to Recover the Missing Data Value for An Attribute in a Database. The 7th World Multiconference on Systemics, Cybernetics and Informatics (2003) 1-6
13. Shigcyuki. Oba, Masa-aki. Sato, Ichiro. Takemasa, Morito. Monden, Ken-ichi. Matsubara and Shin Ishii: Missing Value Estimation Using Mixture of PCAs. International Conference on Artificial Neural Networks. (2002) 492-497
14. Jerzy W. Grzymala-Busse, Ming Hu: A Comparison of Several Approaches to Missing Attribute Values in Data Mining. 2nd International Conference on Rough Sets and Current Trends in Computing (2000) 378-385

Recognizing and Simulating Sketched Logic Circuits

Marcus Liwicki[1] and Lars Knipping[2]

[1] Department of Computer Science, University of Bern
Neubrückstr. 10, CH-3012 Bern, Switzerland
liwicki@iam.unibe.ch
[2] Freie Universität Berlin, Institut für Informatik
Takustr. 9, 14195 Berlin, Germany
knipping@inf.fu-berlin.de

Abstract. This paper presents a system for recognizing sketched logic circuits in real-time and graphically simulating them afterwords. It has been developed for use in university and school education. Circuit gate symbols are recognized using a multilayer perceptron network. The simulation is fully controlled by hand-drawings, and the inputs to circuits can be defined by writing numbers next to them. In addition to the simulation of simple circuits, recursive circuits can also be handled by the system. Furthermore, clock elements can be added for the purpose of synchronization, and circuits can be stored to be reused as sub-circuits, allowing the user to build arbitrary complex configurations. The usability of the system has been tested in a small video-taped laboratory test.

1 Introduction

In this paper a system for recognizing and simulating sketched logic circuits is presented. The tool has been developed for use in university and school teaching. It is used within a lecture recording system based on a chalkboard as user interface, called E-Chalk [4, 5]. With the E-Chalk application a lecturer can do freehand writing and drawing using a pen-based input device. In the context of the presented system, drawings of logic circuits can be recognized on-line, and the circuits can be graphically simulated for the audience.

A number of systems using sketched inputs for other types of applications have been developed recently. For example DENIM [13] allows to build web pages by drawing, SketchySPICE [9] is a simple-circuit CAD-tool, Tahuti [7] is used for creating UML diagrams by sketches, and ASSIST [1] is a sketched-based CAD-tool. For the E-Chalk system mentioned above, applications for animations of algorithms [3], simulating biological and pulse-coded Neural Networks [10, 11], and interpretation of handwritten Python scripts using the Microsoft Handwriting recognizer [10] have been realized in our previous work. Most of them are used in education for visualizing complex processes.

It is also useful to visualize the processes of logic circuits with simulation engines, like it is done in [8, 12]. These systems rely on drag-and-drop interfaces or textual circuit definitions for input. Previous approaches for recognizing

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 588–594, 2005.

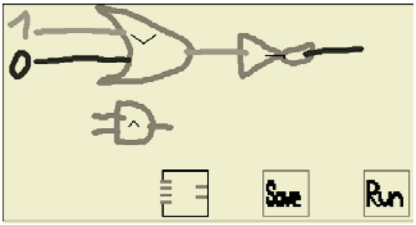


Fig. 1. Screen shot of the circuits simulator chalklet

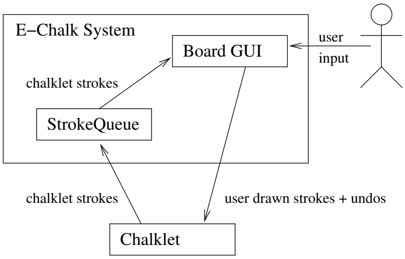


Fig. 2. Illustration of the E-Chalk interface

sketched circuits have also been developed. In [15] a system for off-line recognition focusing on tracing connection lines and their intersections is described. To the knowledge of the authors no integrated system for recognizing sketched logic circuits and simulating them afterwords exists. This would be applicable during lectures with a pen-based interface for education.

The rest of the paper is organized as follows. In Sect. 2 the functionality of the logic circuit simulator is presented, and an overview of the interface to the e-learning system is given. Section 3 describes the recognition for the sketched circuits and Section 4 presents the graph-based simulator. Finally, Sect. 5 draws some conclusions and gives an outlook for future work.

2 User Interface

The system for recognizing and simulating logic circuits provides many functionalities to the user. Several functionalities can be accessed by using special colors or by drawing small strokes within a given region, which is interpreted as activating an element. If the user draws a background colored stroke on a previously sketched element, for instance, it is completely removed by the system. Figure 1 shows a screen shot of the graphical user interface. The system has been produced for processing pen-based on-line data. This data can be acquired with a digitizer tablet, with a digitizing whiteboard, or by using a Tablet PC.

The system is designed as a plug-in, a so called chalklet, for the E-Chalk system [10]. All strokes drawn by the user into the chalklets area are sent to it. The chalklet receives the strokes and reacts for instance by writing strokes into the area to work as a real-time engine. An illustration of the interface to the E-Chalk system is given in Fig. 2. The stroke data consists of a time stamped point sequence, a stroke color, and a drawing width. The color information allows us to define different input modes and the drawing width is used for measuring distances between drawn elements.

The system described in this paper recognizes the five logic symbols for *and*, *or*, *not*, *multiplexer*, and *demultiplexer*. The symbols for the gates follow the notation of the hades system [8], which is also widely used in literature. Examples for these symbols are given in Fig. 3. The user has to start drawing symbols from

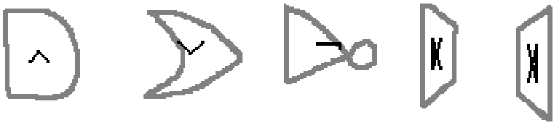


Fig. 3. Gates which are recognized by the system (from left to right): *and*, *or*, *not*, *multiplexer*, and *demultiplexer*; the visual feedback of the system is illustrated in black

one of the ending points at the input side of the gate, and the symbol can be drawn into four directions, i.e. the input side can be on the left, at the top, at the bottom or on the right.

One predefined color is reserved for drawing wires and gates. The system detects the type of the elements and gives a visual feedback at the center of its bounding box. For an example see the five gates in Fig. 3. Each gate can be used with an arbitrary number of inputs and outputs. The leftmost of the three buttons at the bottom of the panel shows the number of inputs and outputs of the circuit, see Fig. 1. For a detailed description of the recognition process see Sect. 3.

For simulating the sketched logic circuit there are two possibilities. First, a simulation can be started by drawing a one (high) or a zero (low) in the color reserved for logic level input. The system then sets the nearest input of the circuit to the corresponding state. The result is visualized by repainting the stroke in the color of the corresponding logic level. As described in Sect. 4 the states of all connected wires are updated immediately. Note that the state of an input can be changed anytime by drawing a new input value. The second possibility for starting the simulation is to activate the button with the caption “Run” shown in Fig. 1, i.e. to draw a small stroke inside the button area. Then all possible combinations of the uninitialized inputs of the circuit are simulated sequentially.

The simulator also allows the user to synchronize the gates with a timer. A clock symbol, drawn in a third reserved color, can be connected to the desired gates of the circuit, see Fig. 4. The connected gate retains its output until one clock cycle elapsed, as if a D-Flip-Flop is connected to the outputs of the gate. A clock can be started and stopped by activating the clock symbol.

Figure 4 also illustrates another feature of the chalklet, namely displaying a state-timing diagram, which can be switched on before starting the application. For every clock cycle the states of the inputs and outputs of the circuit are then illustrated in the diagram. If no timer is present all circuit states are added to the diagram for each change of the circuit input. For clarity the inputs and outputs are numbered from top to bottom in the state-timing diagram, see for example the RS-Flip-Flop shown in Fig. 4 where input I1 is used for resetting and input I2 for setting the output O1.

Another feature is the possibility to save a circuit with a self-defined symbol and to reload it in future sessions. The symbol of the circuit can be drawn in the input and output displaying box mentioned above, see Fig. 4. By activating the button labeled “Save” the graph and the symbol of this circuit is stored into a

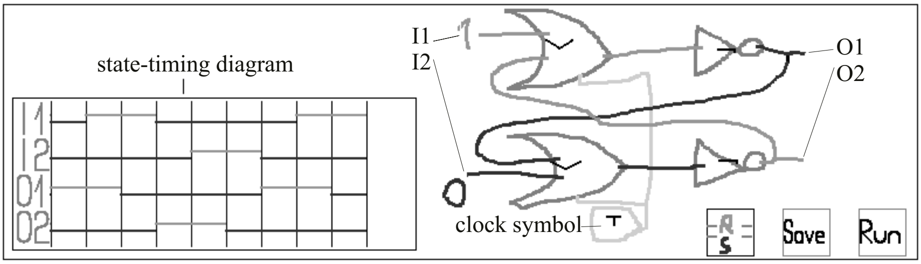


Fig. 4. RS-Flip-Flop and a state-timing diagram

repository of circuits. In later sessions it can be loaded by drawing a rectangle in place of the stored sub-circuit with a fourth reserved color. All symbols of the repository are drawn in the right part of the chalklet area. The user selects the circuit by activating the corresponding symbol box. The system then marks the drawn rectangle with the symbol and the inputs and outputs. See Fig. 5 for an example of integrating a previously defined RS-Flip-Flop into the circuit.

To test the system for usability a small video-taped laboratory test [2] has been conducted. Eight users with different background knowledge about logic circuits received a 15 minute introduction. They were asked to solve exercises and were interviewed afterwards. As a results distance thresholds have been relaxed because some test persons experienced difficulties connecting the elements.

3 Recognition

The recognition of a drawn circuit element is performed whenever a stroke in the color for gate and wire elements is received. First, the system has to determine if a logic symbol or a wire has been drawn. Gates are assumed to be closed, i.e. the distance from the start of the stroke to the end must be below a predefined threshold. Next, the function of the drawn element is detected. In the case of a gate a multilayer perceptron network is applied for classification. In the case of a wire the connections to gates and other wires are identified. Figure 6 illustrates the steps for recognizing a drawn circuit element.

Before the recognizer can be applied to classify the symbols, some preprocessing steps are needed to normalize the input data. Most preprocessing steps are adopted from [6]. First, the stroke is rotated and mirrored so that it starts in the upper left corner and then moves to the right. This normalization procedure reduces the number of possible ways to draw the symbol, one for laying it out with horizontal and one for laying it out with vertical inputs and outputs. The normalization is completed by scaling the stroke to fit into the unit square. Then the data is smoothed by averaging the point coordinates with its previous and next neighbors using the window $(0.25, 0.5, 0.25)$. After that the stroke is interpolated and re-sampled to a fixed number n of equidistant points. Tests on the validation set described below showed $n = 20$ to be a good choice.

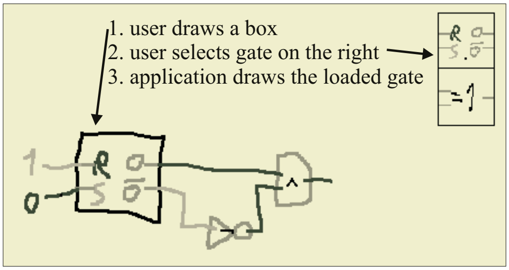

1. user draws a box
2. user selects gate on the right
3. application draws the loaded gate
R
S
Q
Q̄
=1
1
0

Fig. 5. Loading a previously saved RS-Flip-Flop into the chalklet

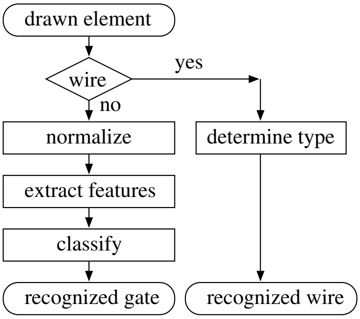


Fig. 6. The recognition process for drawn circuit elements

From the sampled point sequence $s_i = (p_0, p_1, \ldots, p_n)$ the following local features are calculated as input for the classifier: the position (x, y) of each point p_i and the sine and cosine between the line (p_i, p_{i+1}) and the x-axis. In addition to the local features of each point, some global features are introduced. First, to avoid data loss during re-sampling, the coordinates of the points obtained in the first five interpolation steps of the recursive algorithm, are used as features. Next, twelve further points are used. These are the nearest points to the corner points of the bounding box using three different metrics:

$$\delta_\alpha(p, q) = \alpha * |p_x - q_x| + |p_y - q_y| \tag{1}$$

For the first metric α is set to one. For the second metric α is five and for the third metric it is set to 0.2. The described twelve points help to differentiate between the *and* and the *multiplexer* symbol. Additional local features proposed in [6] turned out not to be useful for the recognition of logic symbols in the experiments on the validation set described below. For further details see [14]

The classifier we use is a fully connected multilayer sigmoid perceptron network. The network has a single hidden layer. It has been trained using the backpropagation algorithm RProb. For further informations on the training algorithm see [16].

The training data consists of 700 drawn symbols, i.e. 70 samples for each circuit element and both possible orientations. These symbols have been drawn by five persons. It was assured that symbols representing extreme cases have been added. The data was split into 500 samples for training the network and 200 samples for validating the number of perceptrons in the hidden layer. The optimum was found at eleven elements in the hidden layer.

For wires, three different kinds have to be distinguished: inputs, outputs, and connectors. For classification, it is first analyzed whether it touches the input or output side of a gate. Then any connection via a touching wire is recursively considered.

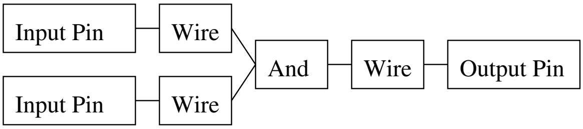


Fig. 7. Graph representation of an *and* with two inputs and one output - the input and output pins are not visible in the GUI

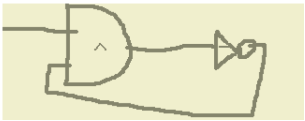


Fig. 8. The output state of the *and* remains undefined because it is not stable

4 Simulation

For the simulation an internal graph representation of the circuit is created. This representation is motivated from the graph representation in [8]. The edges of the graph are the connection points of wires and gates or between two wires. For an example graph representation for an elementary circuit see Fig. 7. For the connection of gates and wires the direction of the signal flow is defined by the direction of the gate symbol. For all other connections the direction is calculated recursively by using the information of the neighboring wires. This direction determines if a wire end serves as input or output of the circuit.

As described in Sect. 2 the engine can simulate the logic circuits asynchronously and time synchronously. The asynchronous simulation starts immediately after each change of the inputs of the circuit. The new signal is transmitted through the wire to all neighboring elements. For each neighbor the new output states of the corresponding elements are computed. If any state changes, the new signal is further transmitted until a stable state is reached. A wire's state is set to undefined if no stable state exists, see for example Fig. 8. If the signal of two symbols is transmitted by a gateless connection to one wire, the connection behaves as an *or* and transmits the new signal to all neighbors. In time synchronous simulation, all changed connections are stored in a queue. At the end of each clock cycle all output changes are transmitted to the corresponding wires.

5 Conclusions and Future Work

The system presented in this paper is able to recognize and simulate sketched logic circuits in real-time. All user interaction can be handled with a pen-input device. The main functionalities are recognizing and simulating wires and gates with an arbitrary number of inputs and outputs, synchronizing the gates with a timer, displaying a state-timing diagram, and saving circuits with user-defined symbols for using them as sub-circuits in a more complex configuration. The system has been developed for use in education.

In the future we plan to enhance the logic circuit recognition system with more functionalities. The number of recognized gates will be enlarged. We also want the user not having to change the color and switch between different modes, to make the creation and simulation of complex circuits even faster. Part of the methods in this work are also suitable in other areas of sketch recognition, such as a molecule recognizer for education in chemistry or an interactive assistant for geometric proving in mathematics.

References

1. Alvarado, C., Davis, R.: Resolving ambiguities to create a natural computer-based sketching environment. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Seattle (WA), USA (2001) 1365–1374
2. Dumas, J.S., Redish, J.C.: A Practical Guide to Usability Testing. Intellect Books (1999)
3. Esponda Argüero, M.: A New Algorithmic Framework for the Classroom and for the Internet. PhD thesis, Freie Universität Berlin, Institut für Informatik (2004)
4. Friedland, G., Knipping, L., Schulte, J., Tapia, E.: E-Chalk: A lecture recording system using the chalkboard metaphor. Interactive Technology and Smart Education (ITSE) **1** (2004) 9–20
5. Friedland, G., Knipping, L., Tapia, E.: Web based lectures produced by AI supported classroom teaching. International Journal of Artificial Intelligence Tools (IJAIT) **13** (2004) 367–382
6. Guerfali, W., Plamondon, R.: Normalizing and restoring on-line handwriting. Pattern Recognition **26** (1993) 419–431
7. Hammond, T., Davis, R.: Tahuti: A geometrical sketch recognition system for uml class diagrams. In Stahovich, T., Landay, J., Davis, R., eds.: Papers from 2002 AAAI Spring Symposium on Sketch Understanding, Palo Alto (CA), USA, AAAI Press (2002) 59–66 Technical Report SS-02-08.
8. Hendrich, N.: HADES: The Hamburg design system. In: EASA'98 (European Academic Software Award)/ Alt-C Conference: Lifelong Learning on a Connected Plane, Oxford, UK (1998)
9. Hong, J.I., Landay, J.A.: SATIN: a toolkit for informal ink-based applications. In: Proceedings of the thirteenth annual ACM symposium on User interface software and technology (UIST), San Diego (CA), USA, ACM press (2000) 63–72
10. Knipping, L.: An Electronic Chalkboard for Classroom and Distance Teaching. PhD thesis, Freie Universität Berlin, Institut für Informatik (2005)
11. Krupina, O.: Client-Server Architecture for a Neural Simulation Tool. PhD thesis, Fachbereich Mathematik und Informatik, Freie Universität Berlin (2005) To appear.
12. Li, L., Huang, H., Tropper, C.: Dvs: An object-oriented framework for distributed verilog simulation. In: 17th Workshop on Parallel and Distributed Simulation. (2003) 173 – 180
13. Lin, J., Newman, M.W., Hong, J.I., Landay, J.A.: DENIM: finding a tighter fit between tools and practice for web site design. In: Proceedings of the Conference on Human Factors in Computing Systems (CHI), The Hague, The Netherlands, ACM press (2000) 510–517
14. Liwicki, M.: Erkennung und Simulation von logischen Schaltungen für E-Chalk. Diploma's thesis, Fachbereich Mathematik und Informatik, Freie Universität Berlin (2004)
15. N. Masayuki, A. Takeshi, I.H.: Pattern recognition for logical circuits diagrams written by freehand. Technical Report 015 - 002, SIGNotes Computer Graphics and cad (1984)
16. Rojas, R.: Neural Networks. A Systematic Introduction. Springer Verlag (1996)

Automatic MLP Weight Regularization on Mineralization Prediction Tasks

Andrew Skabar

Department of Computer Science and Computer Engineering
La Trobe University, Victoria, 3086, Australia
a.skabar@latrobe.edu.au

Abstract. Conventional neural network training methods attempt to find a set of values for the network weights by minimizing an error function using some gradient descent based technique. In order to achieve good generalization performance, it is usually necessary to introduce a regularization term into the error function to prevent weights becoming overly large. In the conventional approach, the regularization coefficient, which controls the degree to which large weights are penalized, must be optimized outside of the weight training procedure, and this is usually done by means of a cross-validation procedure in which some training examples are held out, thereby reducing the number of examples available for weight optimization. Bayesian methods provide a means of optimizing these coefficients within the weight optimization procedure. This paper reports on the application of Bayesian MLP techniques to the task of predicting mineralization potential from geoscientific data. Results demonstrate that the Bayesian approach results in similar maps to the conventional MLP approach, while avoiding the complex cross-validation procedure required by the latter.

1 Introduction

Conventional neural network training methods attempt to find a single set of values for the network weights by minimizing an error function using some gradient descent based technique. The error function is chosen such that the resulting network represents the most probable network, given the data. This is generally referred to as the ML (maximum likelihood) approach. In order to achieve good generalization performance, a term is introduced into the error function to penalize large magnitude weights. The degree to which large weights are penalized depend on the weight decay parameter, whose value must usually be determined using cross-validation.

In contrast to the ML approach, Bayesian MLP methods do not attempt to find a single best weight vector; rather, they attempt to infer the posterior distribution of the weights, given the data. Weight vectors can then be sampled from this distribution, each vector representing a distinct MLP. Given some novel example, each of the sampled networks can then be applied to this example, with the resulting prediction being the average prediction over the collection of networks, weighted by the posterior probability of the network given the training data. Thus, whereas the conventional MLP approach optimizes over parameters, the Bayesian approach integrates over parameters [1]. One of the distinct advantages of the Bayesian approach is that

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 595–601, 2005.

the weight decay parameter can be set automatically; that is, there is no need to use separate training and validation data. A review of Bayesian techniques for MLPs is provided in [2].

Mineral potential mapping is a process whereby a set of input maps, each representing a distinct geo-scientific variable, are combined, using some mapping function, to produce a single map ranking areas according to their potential to host deposits of a particular type [3]. From an inductive learning perspective, the main task is to discover this mapping function, and MLPs have been shown to be suitable [4][5].

Because of the nature of the training data used in mineral potential mapping tasks, determining a suitable value for the regularization coefficient is critical in achieving good generalization performance, and, in the conventional approach, the value of this coefficient must be determined using cross-validation. A novel cross-validation procedure for optimizing the value of the regularization coefficient on mineral potential mapping tasks was described in [6]. However, cross-validation has several disadvantages: (i) it is computationally expensive, as the weight training procedure must be applied many times; (ii) it is noisy in that the final results will depend on the particular data partitions used for cross-validation, and (iii) it makes fewer examples available for weight optimization. This last point is particularly pertinent in the case of mineral potential mapping, because mineralization is a rare event, and the number of examples corresponding to known mineralization is already a very small proportion of the total number of cells in the study area.

This paper describes the application of Bayesian MLP techniques to mapping gold mineralization potential in the Castlemaine region of Victoria. Section 2 provides background into the area of Bayesian techniques for MLPs. Section 3 introduces the mineral potential mapping problem and describes how MLPs can be applied to this task. Section 4 provides empirical results and a discussion. Section 5 concludes the paper.

2 Bayesian Techniques for MLPs

Bayes' Theorem provides a means of determining the posterior distribution of the weights, **w**, given the data, D:

$$p(\mathbf{w} \mid D) = \frac{p(D \mid \mathbf{w})p(\mathbf{w})}{p(D)} \tag{1}$$

where $p(\mathbf{w}|D)$ is the posterior weight distribution, $p(\mathbf{w})$ is the prior weight distribution, and $p(D|\mathbf{w})$ is the likelihood function. The form of $p(D|\mathbf{w})$ depends on the problem, (*i.e.*, regression or classification), and for a binary classification task is given by

$$p(D \mid \mathbf{w}) = \exp\left(-\left(-\sum_{n}\left\{t^{n}\ln y(\mathbf{x}^{n},\mathbf{w}) + (1-t)^{n}\ln(1-y(\mathbf{x}^{n},\mathbf{w}))\right\}\right)\right) \tag{2}$$

where $y(\mathbf{x}^n,\mathbf{w})$ is the network output thresholded using the logistic function; i.e.,

$$y(\mathbf{x}^{n},\mathbf{w}) = h(u) \text{ where } u = \sum_{j=0}^{N_1} w_{kj}\, g\left(\sum_{i=0}^{N_0} w_{ji}x_i^n\right) \tag{3}$$

where N_0 is the number of inputs, N_1 is the number of units in a hidden layer, w_{ji} is the weight connecting input unit i with hidden unit j, w_{kj} is the weight connecting hidden unit *j* with output unit *k*, $h(u) = (1 + \exp(-u))^{-1}$ (*i.e.*, a *sigmoid* function), and $g(u)$ is either a sigmoid, or some other continuous, differentiable, nonlinear function. The prior weight distribution, $p(\mathbf{w})$, is commonly assumed to be Gaussian with mean 0 and inverse variance α:

$$p(\mathbf{w}) = \left(\frac{\alpha}{2\pi}\right)^{m/2} \exp\left(-\frac{\alpha}{2}\sum_{i=1}^{m} w_i^{\,2}\right) \tag{4}$$

where *m* is the number of weights in the network. The rate of weight decay is controlled by the parameter α, which is often referred to as a *hyperparameter* [7]. Thus, neural networks whose weight vectors have small magnitudes are assumed to be more probable than those with large magnitudes.

The Bayesian approach involves integrating the α out of the prediction, i.e.,

$$p(\mathbf{w} \mid D) = \iint p(\mathbf{w} \mid \alpha, \beta) p(\alpha, \beta \mid D) d\alpha d\beta \tag{5}$$

A fully Bayesian treatment would involve defining a prior distribution for α, and then integrating α out. However, because exact integration is analytically intractable, some other approach must be taken.

In this paper we adopt MacKay's *evidence framework* (1992), in which an estimate for α is found by maximizing the likelihood term $p(D|\alpha) = \int p(D|\mathbf{w})p(\mathbf{w}|\alpha)d\mathbf{w}$ (called the *evidence* for α) with respect to α [7]. The use of a Gaussian approximation for $p(\mathbf{w}|D)$ leads to a simple re-estimation formula in which a new value of α is estimated in terms of the eigenvalue/eigenvector decomposition of the Hessian matrix (see [7]).

The practical application of the technique involves the following steps:

1. Choose and initial value for α.
2. Initialize network weights using values drawn from the prior distribution.
3. Train the network to minimize the total error function

$$E_{TOTAL} = \sum_n \left\{ t^n \ln y(\mathbf{x}^n, \mathbf{w}) + (1-t)^n \ln(1 - y(\mathbf{x}^n, \mathbf{w})) \right\} + \frac{\alpha}{2}\sum_{i=1}^{W} w_i^2 \ .$$

4. Re-estimate α using the procedure described in [7] and repeat from Step 3.

Note that in the above description it has been assumed that a single hyperparemeter, α, controls the weight decay. It is, in fact, possible to group weights, and to use a separate hyperparameter for each weight grouping. For example, weights could be grouped into input-to-hidden-layer weights, input-to-hidden-layer biases, hidden-to-output-layer weights, and hidden-to-output-layer biases. These are the weight grouping which have been used in the experiments described in this paper. Alternatively, input-to-hidden-layer weights can be grouped according to the input unit from which they emanate. This latter scheme is used for automatic relevance detection (ARD) [7].

3 MLPs for Mineral Potential Mapping

Mineral potential mapping is the process of producing a map which ranks areas according to their potential to host deposits of a particular type [6]. More formally, the task can be expressed as follows:

Given:

1. Background information provided by m layers of data, each of which represents the value of a distinct geoscientific variable x_i at each pixel *p;*
2. A subset of pixels, each of which is known from historical data to contain one or more deposits of the sought after mineral;

Find:

A function $f(\mathbf{x})$ that assigns to each pixel p in the study area a value that represents the probability that pixel p is mineralized, given the evidence supplied by the background information.

Thus, assuming that the evidence for a pixel p is described by a vector $\mathbf{x} = (x_1, \ldots, x_m)$, the objective is thus to learn a function f: $\mathbf{X} \rightarrow [0,1]$, where $f(\mathbf{x})$ represents the conditional probability that p contains one or more of the known deposits, given the evidence provided by $\mathbf{x}$.

What sets this problem apart from many other problems dealing with spatial prediction is the lack of ground truth information concerning non-deposit cells. For example, in a land classification problem, we would require examples representative of each of the types of land cover, and we would assume that there was sufficient ground truth to be able to identify such examples. However, in the mineralization domain, while we know with certainty that some examples are mineralized, for the greatest majority of cells, we have no information; i.e. they may or may not be mineralized. That is, the absence of a *known* deposit does not mean that the cell is barren. This means that some care must be taken in how the outputs of the MLP are to be interpreted.

Assume the existence of a binary function $g(\mathbf{x})$ that represents the presence or absence of a known deposit in a pixel p with attribute vector $\mathbf{x} = (x_1, x_2, \ldots, v_m)$, where x_k is the value of the k^{th} variable for pixel p, and m is the number of input variables. Thus, if pixel p contains a known occurrence, then $g(\mathbf{x}) = 1$, otherwise $g(\mathbf{x}) = 0$. Now let $f(\mathbf{x})$ be a function whose output is the probability that $g(\mathbf{x}) = 1$. The objective is to learn the function f: $X \rightarrow [0,1]$, such that $f(\mathbf{x}) = P(g(\mathbf{x}) = 1)$. Thus, pixels containing one or more known deposits are assigned a target value of one, and all other pixels in the study area are assigned a target value of zero. Importantly, note that by assigning a pixel a target output of zero, no assumption is being made as to whether or not the pixel contains a deposit. A target output of zero simply means that the pixel does not contain one of the *known* deposits.

The function $f(\mathbf{x})$ can be represented by an MLP. Because the network is required to produce only a single value for each input example, only one output unit is required. Because the output at this unit is to represent a probability, the output of the network should be bounded between 0 and 1, and this can be arranged by using a sigmoidal activation function on the output node.

4 Empirical Results

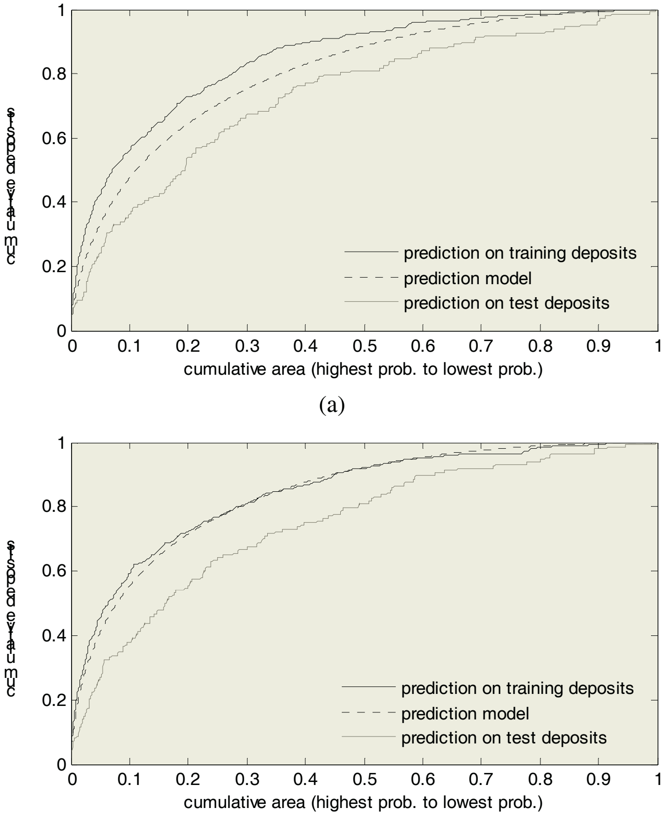
The approach described above has been applied to the production of a mineral potential map showing the favourability for reef gold deposits over a region in the vicinity of the Castlemaine district, Victoria, Australia. Based on a grid-cell resolution of 50m by 50m, the study region was represented by a rectangular grid consisting of 141 cells in the horizontal direction and 206 cells vertically. In total, 16 input layers were used. These included three layers based on magnetics (magnetic field intensity, first derivative of magnetic field intensity, and automatic gain control filtered magnetics); five layers based on radiometrics (Th, U, K, TotalCount, K/Th); seven based on geochemistry (Au, As, Cu, Mo, Pb, W, Zn), and distance to closest fault. The number of documented known reef gold deposits in the study area is 148. The results provided in this section compare the performance of the Bayesian approach described above, with that of conventional MLP training.

In the case of the Bayesian approach, four weight groupings were used: input-to-hidden-layer weights, input-to-hidden-layer biases, hidden-to-output-layer weights, and hidden-to-output-layer bias. The four hyperparameters which control these groups were all initialized to 0.01, but results over several trials indicated that the final results are not sensitive to this initialization. Hyperparemeter re-calculation was performed ten times, with one hundred iterations of the scaled conjugate gradient algorithm [8] performed between hyperparameter updates.

Results for the ML approach are based on the use of a single weight-decay coefficient to control all weights. This was set to a constant value of 2.0. A special cross-validation procedure (see [6]) was used to determine early stopping point and the optimal number of hidden layer units. The cross-validation procedure is based on measuring the likelihood on a set of deposit examples whose class label was hidden during training. Networks with a successively larger number of hidden layer units were trained, and the network with best generalization performance (as determined by holdout deposit examples) was selected. The same cross validation sets were used to monitor training error and halt training when the validation set error began to rise.

Unfortunately it is not possible to adequately reproduce the maps in black and white. However, it is useful to compare the maps on the basis of the favorability *ranking* that they assign to pixels. A convenient means of performing such a comparison is to plot a graph of cumulative deposit frequency versus cumulative area. Such a graph can be constructed by ranking pixels according to their assigned posterior probability value, and plotting the cumulative frequency of deposits (either predicted or observed) against cumulative area as the posterior probability is increased from its minimum to its maximum value. Figure 1 shows the cumulative deposit frequency versus cumulative area curves and Table 1 shows the areas under these curves. The similarity of the two maps can be measured using the Pearson correlation coefficient, which has a value of 0.86, indicating a very strong positive correlation.

From these results, it is clear that the Bayesian MLP approach results are very similar to those obtained using the ML approach. The only significant difference in the results is in the area under the curve for the prediction model, and these can be explained by considering the values of the regularization coefficient in each case. The final values for the regularization coefficients determined using the Bayesian approach are shown in Table 2. In comparison with the constant α of 2.0 used in the ML case, the relatively large value of 32.1 for the input-to-hidden-layer weights in the Bayesian approach ensures that the size of weights in this layer is restricted to small magnitudes, and this is the most likely explanation for the differences observed in the area



(a)

(b)

Fig. 1. Cumulative deposits versus cumulative area represented by pixels ranked from highest probability to lowest probability (a) Bayesian MLP approach; (b) ML approach

Table 1. Area under curves in Figure 1

	Bayesian approach	ML approach
Prediction on training deposits	0.847	0.844
Prediction model	0.804	0.841
Prediction on test deposits	0.734	0.749

Table 2. Regularization coefficient values for weight groupings

Weight Grouping	Regularization coefficient
input-to-hidden-layer weights	32.1
input-to-hidden-layer biases	3.42
hidden-to-output-layer weights	1.73×10^{-2}
hidden-to-output-layer bias	2.68×10^{-2}

under the prediction model curves. The fact that the α value of 2.0 used in the ML approach may have been underestimated suggests that the corresponding map may exhibit some overfitting, but no such overfitting has been observed. The explanation here is that in the ML approach it is the combination of weight regularization and early stopping which prevents overfitting, and both of these have been optimized using cross-validation.

5 Conclusion

In the conventional MLP approach (i.e., what has been referred to above as the ML approach) non-training examples are required to determine the weight-decay coefficients, as well as the early stopping point. In practice, determining optimal values for these parameters is difficult because of interdependencies between them. For example, optimal weight-decay coefficients for a complex model might be different to those for a simpler model. Also, early stopping (if used) may impact on the choice of weight-decay coefficients. These problems are compounded by the fact that cross-validation is noisy in the sense that results will depend to some extent on the particular cross-validation partitions used. The Bayesian approach used in this paper has the very important advantage that it sets the values of the regularization coefficients using the training data alone. This means that all of the data can be used for training, and this is a distinct advantage on mineral potential mapping tasks, in which the number of cells containing known deposits is very small.

References

1. Denison, D.G.T., Holmes, C.C., Mallick, B.K. and Smith, A.F.M., 2002, Bayesian methods for nonlinear classification and regression. John Wiley and Sons, Chichester.
2. Bishop, C., 1995, Neural networks for pattern recognition, Oxford University Press, Oxford.
3. Bonham-Carter, G.F., 1994, Geographic information systems for geoscientists: modeling with GIS, Pergamom Press, Oxford.
4. Brown, W.M., Gedeon,T.D., Groves, D.I. and Barnes, R.G., 2000, Artificial neural networks: a new method for mineral prospectivity mapping: Journal of Australian Earth Sciences, v. 47, no. 4, p.757-770.
5. Skabar, A. 2003 Mineral Potential Mapping using Feed-Forward Neural Networks, *Proceedings of International Joint Conference on Neural Networks*. Portland, Oregon, pp. 1814-1819.
6. Skabar, A., 2004 Optimization of MLP parameters on Mineral Potential Mapping Tasks, *Proceedings of ICOTA: International Conference on Optimization: Techniques and Applications*, 9-11 December 2004, Ballarat, Australia.
7. MacKay, D.J.C. 1992, A practical Bayesian framework for backpropagation networks. In *Neural Computation*, 4(3), pp. 448-472.
8. Møller, M., 1993, A scaled conjugate gradient algorithm for fast supervised learning, in Neural Networks, v. 6, no. 4, p. 523-533.

Integrated Process Modeling for Dynamic B2B Collaboration

Je Yeon Oh[1], Jae-yoon Jung[1], Nam Wook Cho[2], Hoontae Kim[3], and Suk-Ho Kang[1]

[1] Dept. of Industrial Engineering, Seoul National University
San 56-1, Shillim-dong, Kwanak-gu, Seoul, Republic of Korea
{raphael,jyjung,shkang}@ara.snu.ac.kr

[2] Dept. of Industrial and Information System Engineering
Seoul National University of Technology, 172, Gongreung-dong, Nowon-gu, Seoul, Korea
nwcho@snut.ac.kr

[3] Dept. of Industrial and Systems Engineering, Daejin University
San 11-1, Sundan-dong, Poch'on-si, Kyonggi-do, Korea
hoontae@daejin.ac.kr

Abstract. This paper proposes an integrated process modeling methodology for dynamic B2B collaboration. The methodology enables a process architect to effectively design an integrated process in dynamic collaboration environments. An integrated process is composed of internal workflow processes and automated processes, and external collaborative processes with web service interfaces. An integrated process is designed with BPMN, divided into the three types of processes, and automatically transformed to XPDL and BPEL for process enactment. Specifically, collaborative processes are mapped to corresponding actual partners through virtual partners. The concept of virtual partners enables effective maintenance of collaborative processes and flexible exchange of trading partners for B2B process integration in dynamic collaboration environments.

1 Introduction

Many companies have automated their business processes with process enactment systems in order to improve productivity and customer responsiveness. As web service technologies have become widespread to various fields of information systems, business process management systems have also adopted service-oriented architecture (SOA) to promote loosely-coupled B2B integration. A number of researchers have proposed process modeling frameworks by utilizing web service technologies [6, 8, 9]. Specifically, Bussler [2], Jung *et al.*[3], and Leyman *et al.*[4] describe hierarchical structures of B2B process modeling with web services. The previous researches largely focused on how to design external collaborative processes and execute them automatically. However, relatively little attention has been paid to the following: first, most B2B process modeling methodologies have mainly dealt with automatic execution of collaborative processes, not human interaction. Second, previous methodologies did not effectively support flexible exchange of trading partners and configuration management for dynamic B2B collaboration. They did not give

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 602–608, 2005.

careful consideration to process maintenance such as reusability and independence of collaborative processes. To overcome these limitations, an integrated process modeling methodology for dynamic B2B collaboration should support additional requirements as follows:

- Integrated process modeling should cover collaborative processes over multiple organizations and support both automatically executable processes modeling a nd human-interactive process modeling.
- For loosely-coupled B2B integration, automatically executable process models should be seamlessly transformed to web service process standard languages su ch as BPEL.
- To express human initiative tasks, human-interactive process models should su pport human interfaces for process handling and analysis such as object data m odeling, document version management, process simulation, etc.
- For an efficient process design in dynamic collaboration environments, integrat ed process modeling should support a flexible and partner-independent mechani sm of designing collaboration with numerous partners.

To satisfy theses requirements, we present an integrated process modeling methodology for dynamic B2B collaboration. An integrated process model can be divided into three classes: workflow processes, automated processes, and partner-independent collaborative processes. The proposed methodology requires that each type of process be executed by different process engines. The new process modeling methodology enables process designers to effectively design processes that contain human interactive activities, web service activities, and collaborative activities with their partners in dynamic collaboration environments.

2 Integrated Process Modeling

2.1 Basic Concept

Integrated process modeling designs internal and external business processes to accomplish collaboration among multiple companies. The internal process model is composed of workflow processes, automated processes, and collaborative processes. A workflow process means traditional workflow that consists of human interactive activities. An automated process is an automatically executed process without human intervention and it mainly consists of web services. A collaborative process describes the procedure of business-to-business interaction with partners.

First of all, this research extended several components of Business Process Modeling Notation (BPMN) [10] in order to build an integrated process design tool. Process designs following the extended BPMN model are transformed to executable process models in XML Process Definition Language (XPDL) [5] and Business Process Execution Language (BPEL) [1]. All tasks and sub-processes in integrated processes are transformed to XPDL for workflow maintenance and analysis such as process simulation, reengineering, and process enactment by a workflow engine. Automatically executable activities are transformed to BPEL by predefined mapping rules.

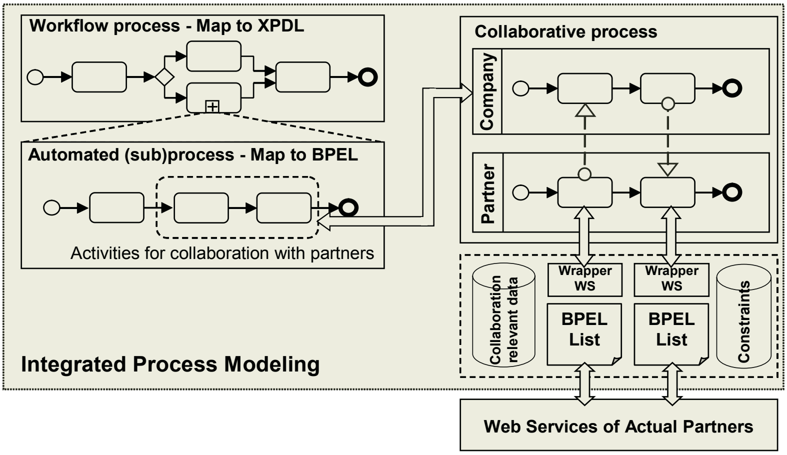

Workflow process - Map to XPDL
Automated (sub)process - Map to BPEL
Activities for collaboration with partners
Integrated Process Modeling
Collaborative process
Company
Partner
Collaboration relevant data
Wrapper WS
Wrapper WS
BPEL List
BPEL List
Constraints
Web Services of Actual Partners

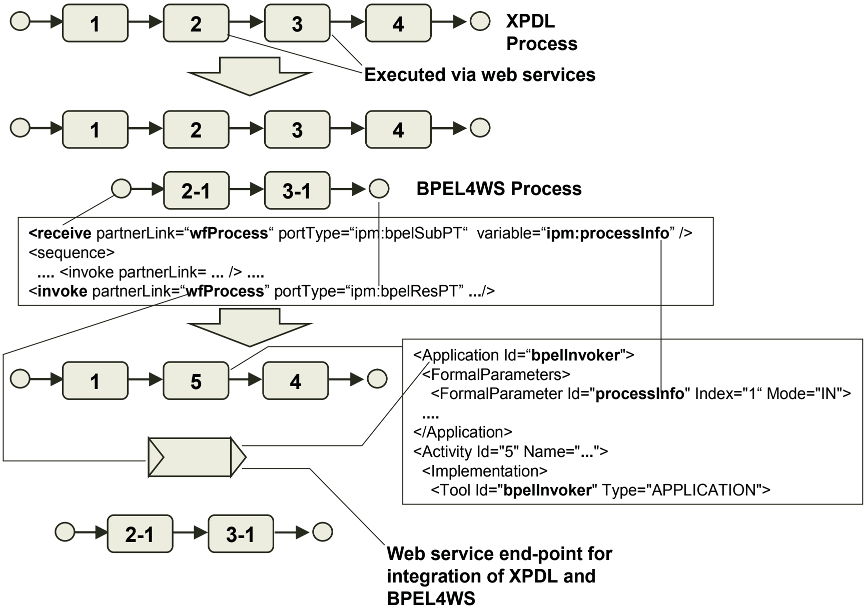
Fig. 1. Integrated process model for dynamic collaboration

Figure 1 shows the concept of the integrated process model for dynamic collaboration. Among the workflow processes and the automated processes, activities designed for B2B collaboration will be generated as new collaborative processes. A collaborative process always consists of a company-side process and a partner-side process. To support partner-independent design of the collaborative process, the concept of virtual partner is provided. Activities of virtual partner-side processes are activated like those of actual partner-side processes and the activities are executed on the internal side of a company. On the other hand, they are interacting with web services of real trading partners outside. Each virtual partner activity is implemented by wrapper web services and maintains a list of BPEL processes that will execute the actual interaction with trading partners. For partner-independent collaboration, a collaboration-relevant data repository and data constraints are also included. When activities of a collaborative process are executed, the wrapper web service of each activity will choose a corresponding BPEL process for its collaboration, instantiates the process, and transform the invocation results of the process into a partner-independent message format.

2.2 Integrated Modeling and Process Deployment

The integrated modeling of workflow processes and automated processes is described in this section. In integrated modeling, a business process is designed without differentiation between workflow processes and automated processes at the design stage. Web services distinguish automated processes from workflow processes before process deployment, and the business process is split into workflow processes and automated processes. The workflow processes and the automated processes are transformed into XPDL and BPEL, respectively.

BPMN provides a user interface for design of an integrated process. The process designed with BPMN is transformed into XPDL. However, since BPMN is not intended for a specific executable process definition language, this research used an extended BPMN in order to provide comprehensive mapping into XPDL.


1
2
3
4
XPDL Process
Executed via web services
1
2
3
4
2-1
3-1
BPEL4WS Process
1
5
4
2-1
3-1
Web service end-point for integration of XPDL and BPEL4WS

```
<receive partnerLink=“wfProcess“ portType=“ipm:bpelSubPT“  variable=“ipm:processInfo” />
<sequence>
  .... <invoke partnerLink= ... /> ....
<invoke partnerLink=“wfProcess” portType=“ipm:bpelResPT” .../>
```

```
<Application Id=“bpelInvoker">
  <FormalParameters>
    <FormalParameter Id="processInfo" Index="1“ Mode="IN">
  ....
</Application>
<Activity Id="5" Name="...">
  <Implementation>
    <Tool Id="bpelInvoker" Type="APPLICATION">
```

Fig. 2. Splitting and transforming progress of workflow processes and automated processes

Extracting automated processes follows mapping of BPMN to XPDL. Automatically executable parts are separated from XPDL as sub-processes and they are transformed into BPEL. The transformation procedure of BPEL consists of two steps: generation of a BPEL process and modification of an XPDL process. The BPEL processes are generated from BPMN models that correspond to an automated sub-process. The generation rule has been developed by the previous research [11]. After this, the XPDL process is modified. An activity is substituted for the part that corresponds to the automated process, and a web service client is assigned to the activity as its application. In this paper, two web service applications are provided to interface between XPDL and BPEL. One is the web service client that initiates the BPEL process by sending a request at execution time. The other is a web service end-point for callback from the BPEL process and a message request from virtual partners, which will be described in Section 2.3.

Figure 2 illustrates the procedure of generating XPDL and BPEL from BPMN. Activity 5 in the Figure is a substituted activity that initiates a BPEL process by using a web service client named “bpelinvoker”. The web service client sends a request message to the BPEL process as partnerLink “wfProcess”. A process instance with information about the variable named “ipm:processInfo” is attached to the request message. The port type “ipm:bpelResPT” is an end point for callback.

2.3 Partner-Independent Collaboration Modeling

In integrated process modeling, a collaborative process consists of a company-side process and a partner-side process. For partner-independent collaboration modeling, a company-side process matches an identical partner-side process, which consists of virtual partner activities. The virtual partner concept eliminates the need of modifying a whole process whenever the company changes its partners. The virtual partner activity receives a request via a web service interface, and then it chooses a corresponding BPEL process in its list. The process is enacted by the BPEL engine and it collaborates with actual partner via external web service call.

The virtual partner activity provides a single interface to each company-side process. This concept resolves the heterogeneity of web services among various partners who perform the same task. There are two types of web services heterogeneity as follows.

- *Heterogeneity of messages* is caused by different definitions of web service message among partners. The discordance can be data heterogeneity or schema heterogeneity. The heterogeneity is similar to that in database integration problem [7]. This kind of heterogeneity can be resolved in two ways. One is the message transformation by XSLT and the other is the transformation logic in BPEL process.

- *Heterogeneity of activity scope* can arise even though the activities of two partner processes perform the same task. In the integrated process modeling, a virtual partner activity can adopt multiple actual partner activities, and the reverse is true (i.e. two types of activities have a relationship with the cardinality (m:n). It can be resolved in three ways, which include 1) the use of BPEL, 2) dummy reply to company-side collaborative activity, dummy-request to real partner activity, and 3) invoking the next virtual partner activity without interaction with the company-side activity.

To resolve the web service heterogeneities, the proposed modeling utilize two repositories, as shown in Figure 1. One is the storage of collaboration relevant data that is used to describe global data. The other is the repository that contains constraints and its invocation addresses.

3 Example Process

This section illustrates the integrated process modeling for dynamic collaboration with an example business process. Figure 3 is a sample process designed with BPMN. All notations in the figure conform to the BPMN specification, where squared boxes are stand-alone systems, rounded boxes are manual or automatic activities, solid arrows are control flows in a system, and dashed arrows are message exchanges among distributed systems. The upper process in the figure shows a workflow process for order handling, which is interacting with an internal system, Customer Information Management System, and an external system, the 3rd Party Warehouse.

The example process is deployed to process definitions in XPDL and BPEL. Figure 4 shows the deployed processes, which include a workflow process, an automated process and a collaborative process. A workflow process is a global process that is enacted, controlled, and analyzed by business administrators. The workflow engine for the process will enact manual tasks directly and invoke automated processes in BPEL to execute automated tasks in legacy information systems of internal organization. In Figure 4, the activity “Initiate Customer and PO Analysis” of the workflow process is nesting an automated process for Sales Management.

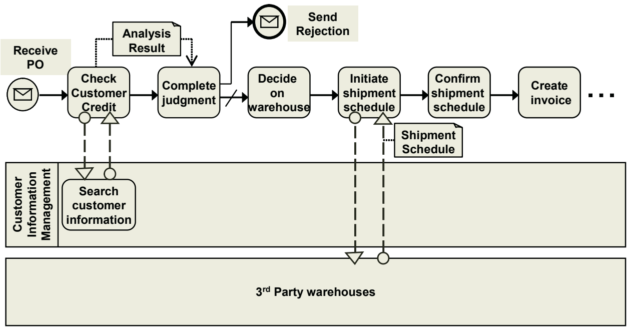


Fig. 3. An example process that represents a process of PO handling

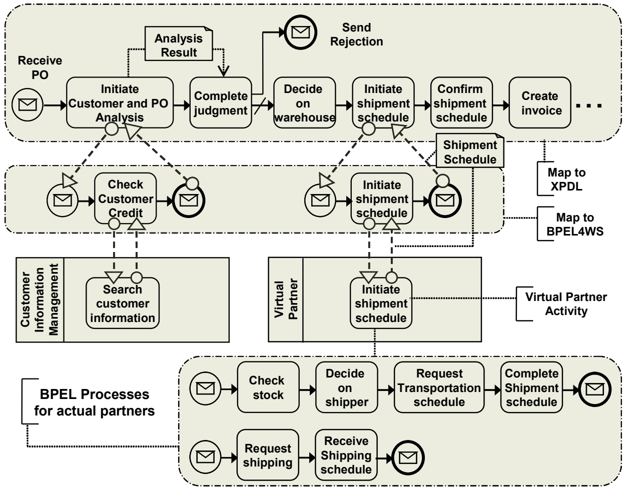

Receive PO
Analysis Result
Send Rejection
Initiate Customer and PO Analysis
Complete judgment
Decide on warehouse
Initiate shipment schedule
Confirm shipment schedule
Create invoice
Shipment Schedule
Check Customer Credit
Initiate shipment schedule
Map to XPDL
Map to BPEL4WS
Customer Information Management
Search customer information
Virtual Partner
Initiate shipment schedule
Virtual Partner Activity
BPEL Processes for actual partners
Check stock
Decide on shipper
Request Transportation schedule
Complete Shipment schedule
Request shipping
Receive Shipping schedule

Fig. 4. Deployed processes of the designed process in Figure 3

In addition, workflow processes can interact with external services of trading partners through collaborative processes, which are defined in BPEL. In Figure 4, the activity "Initiate Shipment Schedule" is interacting with a collaborative process via an activity of virtual partners. Virtual partners are mediators for flexible change of trading partners. The activity of virtual partners in the figure has two collaborative processes of actual partners. If a new partner is added for the "Initiate Shipment Schedule" collaboration, a new BPEL process of the partner will be added to the BPEL list of the virtual activity.

4 Conclusion and Future Work

Current B2B environments require dynamic process-centric collaboration toward loosely-coupled integration. This research proposed an integrated process modeling approach for dynamic collaboration. The approach provides a methodology for designing an integrated process including human-interactive, automated, and collaborative processes. A mechanism for deploying the integrated process to executable processes is also provided. This work enables BPM to satisfy the rapidly increasing need for web service orchestration and dynamic collaboration with numerous business partners by supporting efficient and partner-independent design of business processes.

Challenges remain to be explored. One is the exception handling mechanism. The exception handling procedure that can occur in automated processes may be defined in workflow processes. When exceptions occur in automated processes, they should be forwarded to the workflow process instances, and the exception handling procedure in the workflow process should be initiated and handle the exceptions. The other is selection strategy of a partner in collaborative activities. The measure of the selection can be cost, response time, or other performance metrics.

References

1. Andrews, T. *et al.*: Business Process Execution Language for Web Services (BPEL4WS) Version 1.1. BEA Systems, International Business Machines Corporation, Microsoft Corporation (2003) http://www.ibm.com/developerworks/library/ws-bpel/.
2. Bussler, C.: The application of workflow technology in semantic B2B integration. Distributed and Parallel Databases, vol.12 (2002) 163-191.
3. Jung, J., Hur, W., Kang, S., Kim, H.: Business process choreography for B2B collaboration. IEEE Internet Computing, vol.8, no.1 (2004) 37-45.
4. Leymann, F., Roller, D., Schmidt, M.T.: Web services and business process management. IBM Systems Journal, vol.41, no.2 (2002) 198-211.
5. Norin, R.: Workflow Process Definition Interface - XML Process Definition Language (XPDL). Workflow Management Coalition (WfMC) (2001).
6. Peltz C.: Web services orchestration and choreography. IEEE Computer, vol.36, no.10 (2003) 46-52.
7. Reddy, M.P., Prasad, B.E., Reddy, P.G., Gupta, A.: A methodology for integration of heterogeneous databases. IEEE Transactions on Knowledge and Data Engineering, vol.6, no.6 (1994) 920-933.
8. Schmidt, R.: Enactment of inter-organizational workflows using aspect-element-oriented web services. Proceedings of the 15th International Workshop on Database and Expert Systems Applications (2004) 254-258.
9. Wetzel, I., Klischewski, R.: Serviceflow beyond workflow? IT support for managing inter-organizational service processes. Information Systems, vol.29 (2004) 127-145.
10. White, S.A.: Business Process Modeling Notation (BPMN). Business Process Management Initiative (BPMI) (2004) http://www.bpmn.org/.
11. White, S.A.: XPDL and BPMN. In: L. Fischer, Workflow handbook 2003. Florida: Future Strategies Inc. (2003)

Assessment Methodology on Maturity Level of ISMS

Choon Seong Leem[1], Sangkyun Kim[2], and Hong Joo Lee[3]

[1] Department of Computer and Industrial Engineering, Yonsei University, 134
Shinchondong, Seodaemoongu, Seoul, Korea
leem@yonsei.ac.kr

[2] Somansa, Woorim e-Biz Center, Yangpyeongdong 3-ga, Yeongdeungpogu, Seoul, Korea
saviour@yonsei.ac.kr

[3] Quality & Reliability Lab., DAEWOO Electronics Corp.
412-2 Cheungcheun 2-dong, Bupyeonggu, Incheon, Korea
phileo21@empal.com

Abstract. This paper suggests the evaluation methodology for ISMS (information security management systems) considering technical, managerial, and operational aspects of information security. This methodology includes the evaluation indices, process, and maturity model. We also provide the case study to prove its practical values. This methodology could be used effectively to analyze and evaluate the ISMS of various enterprises.

1 Introduction

As organizations become more and more dependent on their computer-based information systems which play a vital role and important part in their business operations, there must be a greater awareness and concern about the security of these systems. Information security appears on the list of critical success factors of most major organizations today [1,2,3,4,5]. However, to resolve complicated problems of information security, organizations must take not only information systems and technology, but also managerial and operational aspects into consideration [6].

In Korea, according to Information Promotion Law that published in 1995, government agency developed official evaluation systems that provides assessment and certification for information security products. These days, KISA(Korea Information Security Agency) provides official evaluation systems on commercial products of firewall and IDS(Intrusion Detection System) within the country. These kinds of evaluation systems lack in the functionalities to be used in assessment and evaluation for enterprise-wide ISMS because they only focus on technical aspects of package product.

Therefore, enterprises need an evaluation methodology for enterprise-wide ISMS that can assess, analyze, and evaluate their management status of organizational systems on information security.

2 Previous Researches

Previous researches on evaluation systems for information security follows: evaluation systems on security controls, maturity model for information security, and evaluation indices for ISMS.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 609–615, 2005.

TCSEC, ITSEC, and CC are generally used as evaluation systems in information security fields. In early 1970s, US Department of Defense began work on a collection of requirements for trusted computer security for US military. Efforts eventually resulted in the Trusted Computer Security Evaluation Criteria also known as "Orange Book", which was formally published in 1983. The current version of TCSEC was published in 1985. Since then, several other documents have been provided to come up with interpretations of the criteria for networks and databases. Together these books are often referred to as the 'rainbow' series. ITSEC was a common project of the EU members France, Germany, the Netherlands and Great Britain. Unlike TCSEC, ITSEC separates strictly between: functionality and assurance, correctness and effectiveness of assurance. In order to constitute a single, international standard replacing its national ancestors, Canada, France, the Netherlands, Germany and the United States have agreed in 1993 to work on improved evaluation criteria [7]. It's hard to find the previous researches on a maturity model for ISMS. However, NIST's model on security activities and SSE-CMM may be considered as related researches. NIST's model shows security activities in the life cycle of computer system that can be combined with traditional information system methodologies. It consists of initiation, development/acquisition, implementation, operation and maintenance, and disposal [8]. The SSE-CMM is a compilation of the best-known security engineering practices. The SSE-CMM divides security engineering into three basic areas: risk, engineering, and assurance. The SSE-CMM has two dimensions, domain and capability. The domain dimension is perhaps the easier of the two dimensions to understand. This dimension simply consists of all the practices that collectively define security engineering. The capability dimension represents practices that indicate process management and institutionalization capability. These practices are called generic practices as they apply across a wide range of domains. SSE-CMM suggests 5 steps security maturity model: achieved unofficially; planed and managed; defined well; controlled as quantitative; continuous improvement of all [9].

BS7799 and NIST's ASSET model are famous research in evaluation indices for Information security. BS7799 is the most widely recognized security standard in the world. Although it was originally published in the mid-nineties, it was the re-vision of May 1999 which really put it on to the world stage. Ultimately, it evolved into BS EN ISO17799 in December 2000. BS7799(ISO17799) is comprehensive in its coverage of security issues, containing a significant number of control requirements. Compliance with it is consequently a far from trivial task, even for the most security conscious of organizations. It is organized into ten major domains, each covering a different topic or area [10,11,12,13]. BS7799 provides general guidance on the wide variety of topics listed above, but typically does not go into depth. NIST [14] provided a self-assessment guide which is a method for agency officials to determine the current status of their information security programs and, where necessary, establish a target for improvement. It utilizes an extensive questionnaire containing specific control objectives and techniques against which an unclassified system or group of interconnected systems can be tested and measured.

Limits of previous researches are described below. First, previous researches just focus on an evaluation of technical controls. As shown in previous researches, there are three kinds of security controls(management, operational, and technical) to be evaluated. Second, previous researches do not suggest evaluation process and analytic methods on evaluation results. The enterprise just can assess their overall status of ISMS, if the methodology do not supply evaluation process and analytic methods on evaluation results. Third, previous researches do not suggest maturity model. Most enterprises agree that enterprise which accept advanced ISMS can keep superior competitive power. However, there is no solid method for quantitative or qualitative valuation about the maturity model of ISMS.

3 Evaluation Methodology for Organizational ISMS

The evaluation methodology for enterprise-wide ISMS consists of four components. They are evaluation indices, maturity model, decision factors, and process model.

3.1 Evaluation Indices

This paper takes BS7799, NIST's ASSET, and Kim's evaluation indices on information systems into consideration to derive evaluation indices for ISMS [15]. The evaluation indices consist of four levels. The level 1 of evaluation indices consists of a plan level, environment level, support level, and technological level. The level 2 consists of a policy, plan, environment, human resource, support organization, support activity, system operation, and system skill. Figure 1 shows the evaluation indices for ISMS.

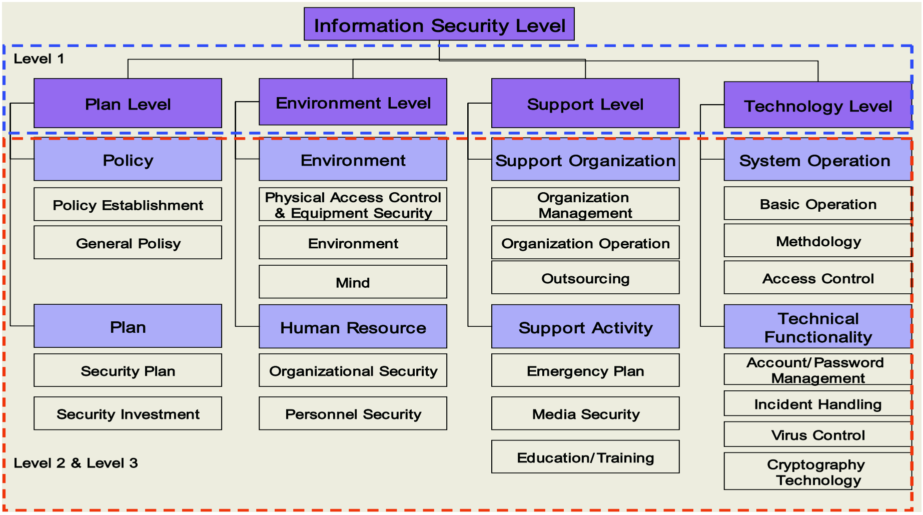


Fig. 1. Evaluation indices for information security management systems

3.2 Maturity Model

This paper provides maturity model of five stages for ISMS including plan level, environment level, support level, and technological level. While SSE-CMM provides maturity levels with process viewpoint of security engineering, this paper suggests maturity levels with synthetic viewpoint of ISMS. The definitions of maturity levels are as follows.

Strategic: The operation of ISMS creates new business opportunities and value of enterprises.

Management - Active: Best practices on ISMS are reviewed periodically. Due diligence and compliance monitoring are well performed.

Management - Passive: Separation of duty and positioning of security officer, administrator, and auditor are completed. Information security strategic planning is established.

Technical: The technical controls for perimeter security are introduced. However, the security policy, procedures, and guidelines are not established.

Functional: Owners of information and systems perform security controls including screen saver, anti-virus vaccine, password-enabled compression utility, and so on by themselves.

3.3 Decision Factors for Maturity Model

The decision factors that describe the characteristics of each maturity level are required to judge the maturity level from evaluation results. The decision factors for maturity model are described in table 1.

3.4 Evaluation Process

The evaluation methodology for ISMS consists of evaluation indices and evaluation process. The evaluation indices consist of the measurement factors and the influence factors. The measurement factors take assessment process to generate present level of ISMS. The influence factors take analysis and management process to generate various kinds of indicator and information that describe special and detailed status of ISMS with dynamic viewpoints. The evaluation process consists of preparation, assessment, analysis, management, and feedback process. This paper takes GAO(United States General Accounting Office), GSA(General Services Administration), and Kim's evaluation process on information systems into consideration to design evaluation process on ISMS[15,16]. The evaluation processes are as follows.

1) Preparation: project initiation; scheduling and resource planning for evaluation.

2) Assessment: data gathering; qualitative or quantitative assessment.

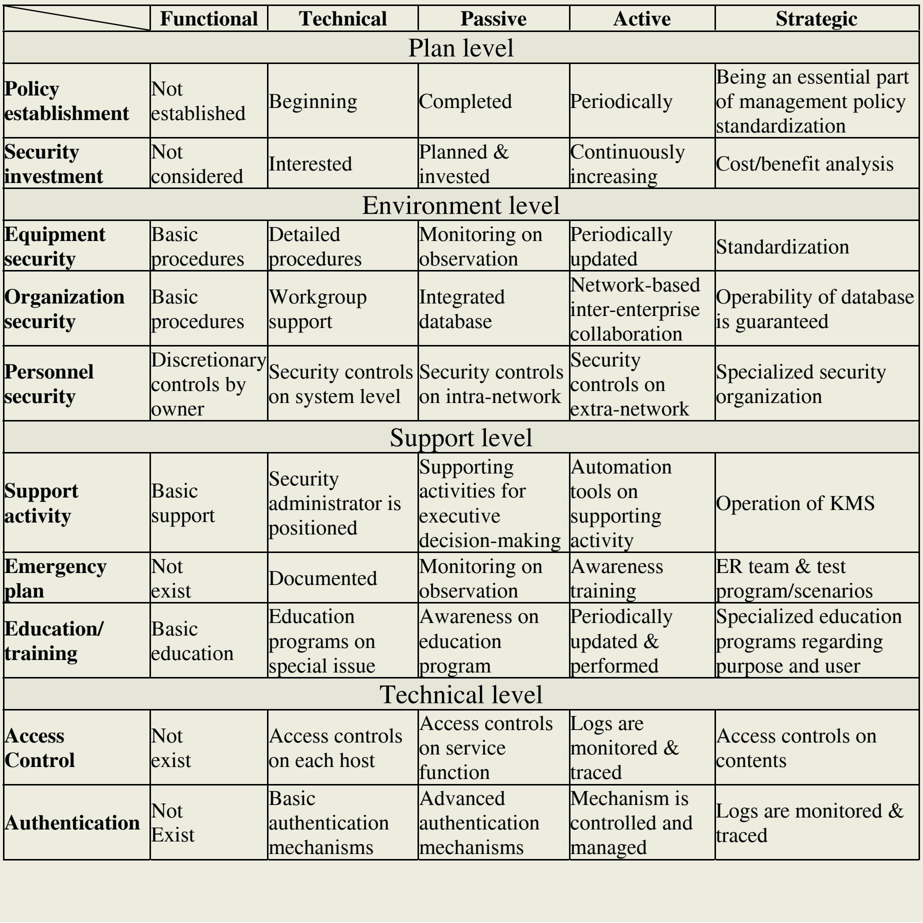
3) Analysis: analysis on assessment results with various viewpoints; reporting.

4) Management: reviews on maturity model; knowledge management; feedback.

4 Verification of Evaluation Indices

We complemented and rectified the evaluation system through a pilot test to determine evaluation items and verify the validity and reliability of the developed evaluation system. The Evaluation indices consist of 4 domains of Plan level, Environment

Table 1. Decision factors for maturity model



	Functional	Technical	Passive	Active	Strategic
Plan level					
Policy establishment	Not established	Beginning	Completed	Periodically	Being an essential part of management policy standardization
Security investment	Not considered	Interested	Planned & invested	Continuously increasing	Cost/benefit analysis
Environment level					
Equipment security	Basic procedures	Detailed procedures	Monitoring on observation	Periodically updated	Standardization
Organization security	Basic procedures	Workgroup support	Integrated database	Network-based inter-enterprise collaboration	Operability of database is guaranteed
Personnel security	Discretionary controls by owner	Security controls on system level	Security controls on intra-network	Security controls on extra-network	Specialized security organization
Support level					
Support activity	Basic support	Security administrator is positioned	Supporting activities for executive decision-making	Automation tools on supporting activity	Operation of KMS
Emergency plan	Not exist	Documented	Monitoring on observation	Awareness training	ER team & test program/scenarios
Education/ training	Basic education	Education programs on special issue	Awareness on education program	Periodically updated & performed	Specialized education programs regarding purpose and user
Technical level					
Access Control	Not exist	Access controls on each host	Access controls on service function	Logs are monitored & traced	Access controls on contents
Authentication	Not Exist	Basic authentication mechanisms	Advanced authentication mechanisms	Mechanism is controlled and managed	Logs are monitored & traced

level, Support level, and Technology level. We verified the validity and reliability between each evaluation domain and evaluation item by factor analysis and reliability analysis. We used five points Likert-type scale, where 1: very inadequate; 2: inadequate; 3: moderate; 4: adequate; 5: very adequate. The survey method includes interviews and questionnaires. The survey was conducted for three weeks.

A sample of 100 usable responses was obtained from various sources. The industries represented in the sample were communication and services (50%), and information consulting and system implementation services (50%). The respondents identified themselves as top manager (8%), middle manager (34%), and worker (58%). The respondents identified their job as strategy planning (46%), development/ maintenance (46%), business operation (2%), and consulting/ education (6%). The respondents had 5 years experience on average in their job and all respondents had college or university degrees in management and economics (10%), and engineering (90%). The validity and reliability of the developed model was proved by factor analysis and reliability analysis. We used the SPSS instrument. From the analysis results on the evaluation items with Factor Loading over average 0.5 and Crombach's Alpha over average 0.6, we can judge that the developed evaluation model has considerable validity and reliability.

5 Case Study

In this case study, AHP was applied to an evaluation project in which ABC Co. Ltd. wanted to analyze their ISMS. The process model described in section 3.4 was used. In this case study, AHP was applied to generate weight factors of evaluation indices. The judgments were elicited from the security experts in the security solution provider and government agency. Table 2 shows weight factors of evaluation indices.

Table 2. Weight factors of evaluation indices

Level 1	Plan level		Environment level		Support level		Technology level	
Weight	0.28		0.1		0.18		0.44	
Level 2	Policy	Plan	Environment	Human resource	Support organization	Support activity	System operation	Technical functionality
Weight	0.31	0.69	0.44	0.56	0.53	0.47	0.59	0.41

This company ranked "active" level of ISMS with 65.7 point. The best practices on ISMS are reviewing it periodically. A due diligence and compliance monitoring are well performed as planned. A plan level is 61.93 point: security policy is periodically updated; investment amount on information security is continuously increasing. An environment level is 77.35 point: mandatory controls on equipment are performing and updating it periodically; network-based inter-enterprise collaboration; security controls on extra-network. A support level is 50.39 point: supporting activities for executive decision-making are performed; importance of education program on information security is well defined. Technical level is 71.73 point: access control logs are gathered, monitored, and traced; authentication mechanism is controlled and managed.

6 Conclusion

This paper suggests an evaluation methodology which consists of evaluation indices, evaluation process, and maturity model. This methodology could be applied to assess, analyze, and manage the current and on-going status of enterprise-wide ISMS. The maturity model may indicate the enterprise's competitive values on information security fields against their competitor or business partners. This methodology supports various scales of enterprise from mid-size to large-size with objective and time-effective manners. The analytic results on maturity level provide how enterprises could upgrade their current status of ISMS. The accumulated information on evaluation results of various enterprises may suggest how government agencies should drive a national policy on ISMS.

Acknowledgments

This work is supported by Korea Information Security Agency, in Korea.

References

1. Kim, S. and Leem, C.S.: Security of the Internet-based Instant Messenger: Risks and Safeguards. Internet Research: Electronic Networking Applications and Policy, Vol.15, No.1, Emerald (2005)
2. Kim, S., Lee, H.J. and Leem, C.S.: Applying the ISO17799 Baseline Controls as a Security Engineering Principle under the Sarbanes-Oxley Act. Lecture Series on Computer Science and Computational Sciences, Vol.1, VSP International Science Publishers (2004)
3. Kim, S. and Leem, C.S.: Implementation of the Security System for Instant Messengers. Lecture Notes in Computer Science, Vol.3314, Springer Verlag (2004)
4. Kim, S. and Leem, C.S.: An Information Engineering Methodology for the Security Strategy Planning. Lecture Notes in Computer Science, Vol.3043, Springer Verlag (2004)
5. Kang, J.B.: Internet Revolution and Internet Security. Triangle press (2001)
6. Shin, D.J: Internet Information Security. Dongil Press (2001)
7. CC Project Team: Common criteria for Information Technology Security Evaluation. Common Criteria Project (1998)
8. NIST: An Introduction to Computer Security : The NIST Handbook. National Institute of Standards and Technology (1995)
9. SSE-CMM Project Team: Systems Security Engineering Capability Maturity Model. SEI of CMU (1999)
10. BSI: BS7799. BSI (1999)
11. Kim, J.D. and Na, K.S.: Measuring of Index of Information Security by Vulnerability Estimation - Information Property Value Weight. Information Security and Cryptology (2000)
12. Barnard, L.: The Evaluation and Certification of Information Security Against BS7799. Information Management & Computer Security, Vol.6, No.2 (1998)
13. Solms, R.V.: Information Security Management: the Code of Practice for Information Security Management(BS7799). Information Management & Computer Security, Vol.6, No. 2 (1998)
14. NIST: Security Assessment Guide Information Technology Systems. National Institute of Standards and Technology (2001)
15. Kim, I.J., Leem, C.S.: Development and Implementation of an Integrated Evaluation System for Continuous Maturity of IS Performance. Journal of the Korean Institute of Industrial Engineering, Vol.29, No.1 (2003)
16. GAO: Executive Guide - Measuring Performance and Demonstrating Results of Information Technology Investments. GAO (1998)

CSFs for HCI in Ubiquitous Computing Environments

Hong Joo Lee[1], Sangkyun Kim[2], and Choon Seong Leem[3]

[1] Quality & Reliability Lab., DAEWOO Electronics Corp.
412-2 Cheungcheun 2-dong, Bupyeonggu, Incheon, Korea
phileo21@empas.com

[2] Somansa, Woorim e-Biz Center, Yangpyeongdong 3-ga, Yeongdeungpogu, Seoul, Korea
saviour@yonsei.ac.kr

[3] Department of Computer and Industrial Engineering, Yonsei University
134, Shinchondong, Seodaemoongu, Seoul 129-749, Korea
leem@yonsei.ac.kr

Abstract. The vacuum cleaning robot provides various services in home environments. The latest robots provide the functionalities of cleaning up a house automatically and patrolling for home security using camera systems. In this paper, we suggest the user-friendly interfaces for vacuum cleaning robot with the criteria for successful HCI(Human Computer Interaction) of Jakob Nielsen.

1 Introduction

The 21st century is an era of the New Technology. From day to day, we are adopting these new technologies to improve the quality of human life. The vacuum cleaning robot is one of the most important research issue between various new technologies. The robot for home service assists the housework of house keeper. Thus, a development of user-friendly interfaces of this robot is very important and should not be ignored[1]. In this paper, we provide the CSF(Critical Success Factor)s for user-friendly interfaces in ubiquitous computing environments.

2 Previous Researches

The HCI stands for human computer interaction[2]. In computer science perspectives, HCI deals with the interactions between one or more humans and one or more computers. According to Sjoerd Michels[2], HCI can be defined as: "the part of a computer program responsible for establishing the common ground with a particular (i.e. well known) user. His task is accomplished by expanding and maintaining this common ground throughout the interaction process with the application. Whenever possible, direct manipulation of familiar objects should be the leading interaction principle." This definition mentions the direct manipulation of familiar objects. The direct manipulation is more possible if the objects are known from the real world or from other HCIs. The user is more likely to trust an object that is familiar. This definition also hints at the goal of HCI, which is to facilitate the interaction between the user

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 616–620, 2005.

and computer. A well designed interface increases a productivity and reduces errors[3]. The computer can be defined as a traditional home or office personal computer or any workstation. The purpose of HCI is to enhance the 'user- friendliness' of a system. This purpose is sometimes wrongfully perceived as opposing the goals which secure a system[4]. For example, a confidentiality of information is desired in a secure system and is accomplished to certain degrees by the use of passwords. However, most users do not remember a long and complex password, which means they will write it down, leading to the potential breakdown of the security of a system. When it comes to usability principles, the fewer the passwords and the simpler the passwords are the better. This appears to highlight a contradiction between security and usability.

3 Key Characteristics of the Robot

There are great anticipations for the application of robots in a wide range of fields such as medical services, social services, housework, and so on. Societal expectations that the robots should contribute to enrich human society are growing continuously.

3.1 Key Features of Latest Robot Systems

The latest robot is consists of motor-driven wheels which move forward, backward, left and right, and rotate independently, CPU which controls the entire robot, and visual processing systems comprised of digital signal processors and custom hardware. The key features of latest robot are described in below.

1) Ability to move autonomously to a designated location using sensor while carefully avoiding obstacles

The robot is capable of perceiving people or things quickly in its surrounding areas while simultaneously measuring its location. This capability is accomplished with the use of cameras. The robot selects cameras by itself from available cameras as necessary which adopts the latest visual processing systems. The visual processing enables the robot to detect and avoid obstacles so that it can move safely to a designated location, thereby making it capable of completing tasks alongside people.

2) Ability to perceive, using sensor and camera

By utilizing the visual processing, the robot can hold things and press buttons such as elevator buttons. The robot can move naturally and smoothly with the use of a central pattern generator which simulates the nervous systems of human beings.

3) Functions to be executed according to the users' needs and to present the information using the most appropriate method

The robot can connect to the Internet, execute users' requests, and present the information in best-fitted methods such as announcing, displaying on its own display, pointing to a separate source, and so on. Various kinds of Internet services can be provided through the robot by using its network-related functions. Self-equipped with

Web server, the robot can be instructed, programmed and remote-controlled from external devices such as a computer, mobile phone, and PDA(Personal Digital Assistants) which do not feature specific applications for the robot.

4) Self-charging
When it is low on power, the robot autonomously moves to a charger to self-charge by using high-density non-contact charging based on an induction charging method to enable safe charging.

5) Ability to maneuver itself flexibly within compact spaces, on slopes, over minor surface gradations, and complete tasks
The robot can move on the spot by using two driving wheels that move independently. It also can move flexibly and smoothly over slopes and uneven surfaces, by utilizing the front and back driving wheels.

3.2 Characteristics of User's Behavior

The characteristics of individual user mean that users has characteristics of resident in their mind[5]. These characteristics consist of four types as described in table 1.

Table 1. Characteristics of user's behavior

Types	Characteristics
Personality characteristics	People have personality characteristics between me and someone
Physical characteristics	Users have physical problems and age characteristics
Cultural characteristics	Personnel of organization have their value and confidence
Motivational characteristics	People have response and attitude about something new

Considering these factors described in table 1, we suggest CSFs for HCI in development of vacuum cleaning robot.

4 CSFs for HCI

In 1983, Apple Computers released the Apple Lisa to the public[6]. The Lisa was one of the first commercially available computers which provide a graphical user interface. The introduction of graphical user interfaces has made the operation of computers much easier and has also led to huge growth in research in the field of HCI. This introduction has led to a number of principles being established[7, 8].

One of the key players in the field of HCI is Jakob Nielsen. He has been involved in HCI and usability for many years and has developed a list of ten criteria for a successful HCI[9]. Based on his research, we suggest CSFs for HCI.

Table 2. CSFs for HCI

CSF	Define
Visibility of system status	It is important for the user to be able to observe the internal state of the system.
Match between system and the real world	An HCI which uses real-world metaphors is easier to learn and understand. The user operates a system easily using icon. (an icon is a picture which represents a particular system function)
User control and freedom	If the users press a button by mistake, they will need a clearly marked exit path.
Consistency and standards	Operating buttons, situations and actions need to be consistent and have the same meaning.
Error prevention	It is obviously best to prevent errors in the first time through careful design. However, errors do occur and they need to be handled in the best possible way.
Recognition rather than recall	The user should not have to remember one session to another. Rather, the user should be able to 'recognize' what is happening.
Flexibility and efficiency of use	The system should be efficient and flexible to use. Reasonable system design and operation raise a productivity.
Help and documentation	Help functionality needs to be context-sensitive and easy to search.

5 Conclusion

In these days, the newly developed products breathe life into human beings with new technology. For the preparation of emergence of ubiquitous computing environments, we present our study on the vacuum cleaning robot which is one of the famous product based on new technologies, and improve the quality of life. In this paper, based on the characteristics of user behavior, we suggest CSFs for HCI to provide user-friendly operation of vacuum cleaning robot in ubiquitous computing environments.

References

1. Lee, H.J. and Lee, J.W.: Ubiquitous Innovation, E-co book (2004)
2. Michels, S.: Look and Feel!, Masters Thesis. Tilburg university (1995)
3. Schneiderman, B.: Sparks of Innovation in Human-Computer Interaction. Human-Computer Interaction Laboratory (1993)
4. Botha, R.A., Principal Lecturer, Business Information Systems, Faculty of Computer Studies, Port Elizabeth Technikon, South Africa (2002)
5. Kim, J.W.: Introduction to Human Computer Interaction. A Graphics (2005)
6. Myers, B.A.: A Brief History of Human Computer Interaction Technology. ACM Interactions, Vol. 5 (1998) 44–54.
7. Carroll, J. (ed.): HCI Models, Theories, & Frameworks: Toward a Multidisciplinary Science. Morgan Kaufmann (2003)
8. Nielsen, J.: Usability Engineering. Academic Press Inc (1994)
9. Nielsen, J.: Ten Usability Heuristics, on the Web site http://www.useit.com/papers/heuristic/heuristic_list.html (accessed in 2005)

Practical Design Recovery Techniques for Embedded Operating System on Complying with RTCA/DO-178B and ISO/IEC15408

Minhyung Kim[1], Sangkyun Kim[2], and Myungwhan Choi[3]

[1] Department of Computer and Industrial Engineering, Yonsei University, 134
Shinchondong, Seodaemoongu, Seoul, Korea
mentor@yonsei.ac.kr

[2] Somansa, Woorim e-Biz center, yangpyeongdong 3-ga
Yeongdeungpogu, Seoul, Korea
saviour@yonsei.ac.kr

[3] Department of Computer Science, AS1007,Sogang University, #1
Shinsoodong, Mapogu, Seoul, Korea
mchoi@ccs.sogang.ac.kr

Abstract. As robustness, stability, and security have been emphasized as the critical success factors of software and systems in the various fields of industry, achieving certification based on a verification of compliance with standards is regarded as the best solution which proves a reliability of product and provides a great marketing tool. Despite such an importance, most developers have scruple about documenting development processes sufficiently. Furthermore, most of software companies try to get a certification after a development of product. As a result, applicants preparing certification are in a quandary about making sufficient materials for certification process. To solve this problem, we present the practical techniques which could be used to recover the details of software design from product and source code in compliance with RTCA/DO-178B and ISO/IEC15408 standards.

1 Introduction

Recently, most of COTS(Commercial-Off-The-Shelf) software companies are willing to pay for the certification project to approve a reliability of their products especially in information security and airborne system area. In addition, many countries and international authorities force to observe the standards(ex. ISO/IEC 15408 in information security and RTCA/DO-178B in civil aircraft) by laws or their regulation [1]. In this situation, a certification means not only a superiority of software product, but also prerequisite factor which sustains a continuity of their business.

However, many companies face with several snags at project preparation on complying with standard. Immaturity of software companies about controlling and assuring software quality is a major problem. For instance, while RTCA/DO-18B and ISO/IEC 15408 standards require evaluation materials which contain entire data from software life cycle for whole development process, most developers have scruple about documenting their developing process or generate these documents insuffi-

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 621–627, 2005.

ciently [2, 3]. So, applicants who prepare certification meet deep trouble if a certification project starts out after completing a development of product due to the lack or absence of data from software life cycle [4]. Another reason which causes this problem is that many companies often try to get a certification after completion of development and packaging system for widening their market.

Limits of previous researches on RTCA/DO-178B and ISO/IEC 15408 are summarized as follows: It's focusing on the summarization of standards; It only analyzes a relevance of another capacity model or development methodologies; It's insufficient to deliver practical techniques to applicants who prepare a certification process.

To solve these problems of previous researches, in this paper, we suggest practical methods to recover the details of software design which are not originally exist, insufficient or missing from source code and product. The method provided in this paper is based on our actual experiences. It proves practical values of our method that we have successfully finished some certification projects with the method of this paper in real world.

2 What Is an Importance of Design Recovery?

As mentioned above, due to the absence or lack of data from software life cycle, many certification projects face with serious problem about the preparation of certification materials. Most applicants try to make these certification materials with reverse engineering instead of rebuilding software or re-engineering software development processes on compliance. However, it's hard to find a referable model or methodology which provides well-defined process yet. It's a reason of this problem that many previous researches about RTCA/DO-178B and ISO/IEC 15408 have been placed too much emphasis on comparability with existing standards or methodologies such as CMM, SSE-CMM, ISO/IEC 12207, ISO/IEC 12119 and etc.

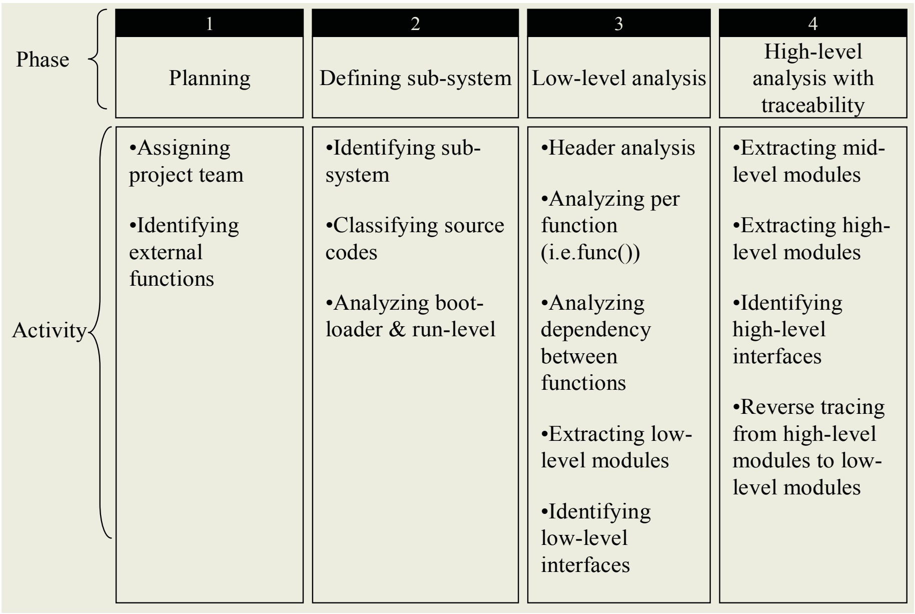
Therefore, it is a key question that what process is useful and referable as practical techniques. In this paper, we propose a practical approach which delivers reverse engineering techniques and provide a framework and subordinate activities.

3 Design Recovery with Defined Processes

Since the past years, we have been conducted two certification projects. Based on our cases, we present a conceptual chart which graphically summarizes the analytical framework of our approach as illustrated in fig. 1. It consists of 4 domain and 14 sub-processes.

3.1 Framework

Above mentioned techniques could be recapitulated as follows: first, planning and extracting a rough design structure with composition of formal(external) functions; second, defining sub-systems and classifying source; third, low-level analysis; and fourth, high-level analysis with traceability.



	1	2	3	4
Phase	Planning	Defining sub-system	Low-level analysis	High-level analysis with traceability
Activity	•Assigning project team •Identifying external functions	•Identifying sub-system •Classifying source codes •Analyzing boot-loader & run-level	•Header analysis •Analyzing per function (i.e.func()) •Analyzing dependency between functions •Extracting low-level modules •Identifying low-level interfaces	•Extracting mid-level modules •Extracting high-level modules •Identifying high-level interfaces •Reverse tracing from high-level modules to low-level modules

Fig. 1. Framework for design recovery

3.2 Subordinate Process

Planning: At planning phase, it lays a strategic and operational plan including team assignment and recognition of external functions from target of evaluation(TOE). The key characteristics of this phase are described in table 1.

Table 1. Characteristics of first phase

Activity	Description
Assigning project team	Organizing project members including architecture experts and entity analyzers.
Identifying external functions	Identifying external functions which mostly described in product's function. (ex. Read & write file, etc.)

Till the completion of certification project, entity analyzers mostly concentrate upon an analysis of function units from source code. At the same time, architecture experts play an important role in analysis of relationships between function units and extraction of mid or high level modules with interfaces. However, external functions for operating system primarily function as an interfacing with human and networking with other systems. It may be possible to get referable information from NIST Common Criteria evaluation and validation scheme [5].

Defining Sub-system: To define sub-system is most important process in design recovery and certification project. It brings out from composition of functions which

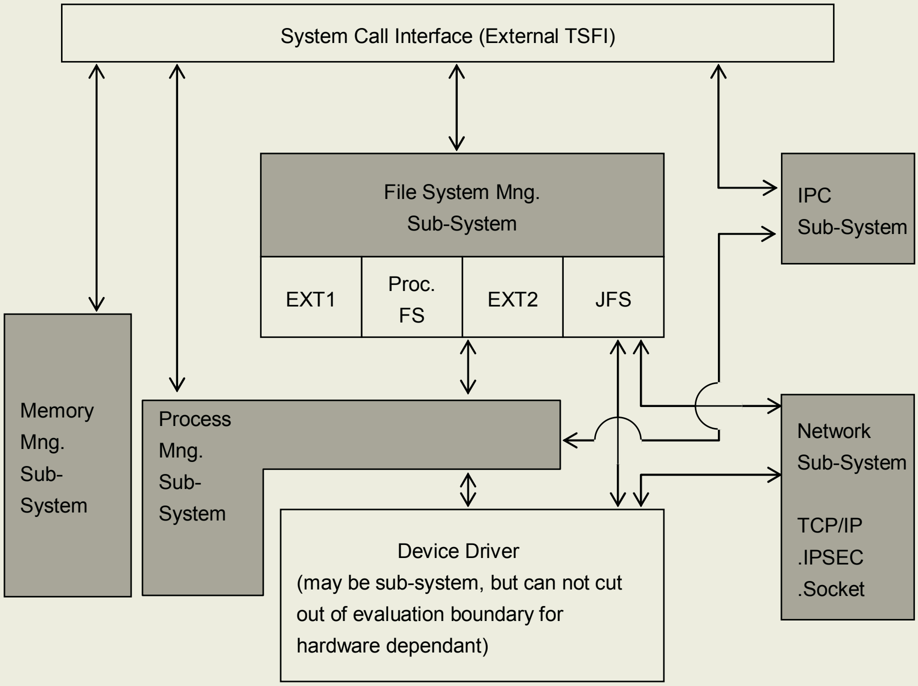


Fig. 2. Identification of sub-system

Table 2. Characteristics of second phas

Activity	Description
Identifying sub-systems	Identifying sub-systems based on composition of external functions. (ex. Process, file-system, memory management, IPC, network). Device drivers and other platform-dependant codes could be excluded from evaluation scope.
Classifying source code	Classifying source code according to type of sub-systems.
Analyzing boot loader and run level	It may be helpful to explain operating environments and initializing procedures.

cannot be combined any more. Usually, an operating system includes Real-Time OS(RTOS) and embedded OS which provides four or five root functions to deserve as sub-system.

The key characteristics of this phase are described in table 2.

Low-Level Analysis: At low-level analysis phase, it includes an analysis of function units from source code (i.e. func() source code analysis) and defines low-level modules from functions. A low-level module is a set of func(), function of lowest level from source code, which supports upper functions. Interfaces between low-level modules are also recognized at this time. The key characteristics of this phase are described in table 3.

Table 3. Characteristics of third phase

Activity	Description
Header analysis	Header contains fundamental information about structures, macros and so on. Thus, it may be a clue which explains an architecture of modules.
Analysis per function	Analyzing a function name, path, process logic, and input/output of each function in source code.
Analyzing dependency between functions	Comprehending a reference and cross-reference information between functions.
Extracting low-level modules	Merging func(), function of lowest level from source-code, distinguished from other functions. (ex. ping, copy, traceroute and etc.)
Identifying low-level interfaces	An identification of interfaces between low-level modules is derived from function call relationship. It is also illustrated in Fig. 3.

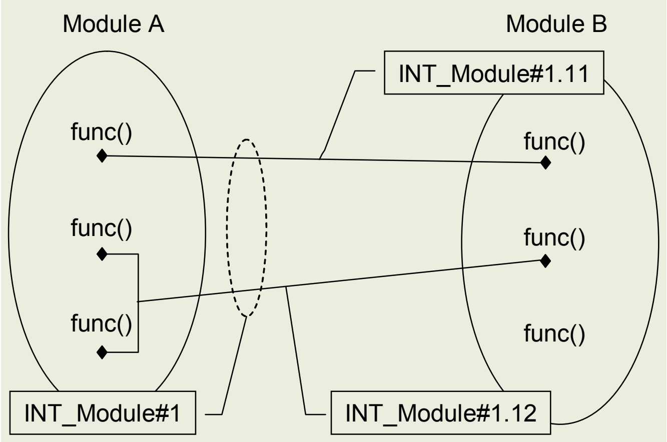


Fig. 3. Method to extract low-level interfaces

High-Level Analysis with Traceability: Finally, in high-level analysis with traceability phase, it extracts mid-level modules and high-level modules on basis of the low-level modules and interfaces. All external functions that we identified at Planning phase must be included in High-level module's functions. The key characteristics of this phase are described in table 4.

Table 4. Task description

Activity	Description
Extracting mid-level modules	Mid-level modules are a set of low-level modules. It must be substantiated in things which could be compiled as a single object(ex. daemon, executable binary and etc.)
Extracting high-level modules	High-level module's functions must contain every external functions that we identified at Planning phase.
Identifying high-level interfaces	High-level interfaces chiefly provide functions which communicate or interact with other external objects. (ex. csh(C shell), ash, ntp daemon, and etc.)
Reverse tracing from high-level modules to low-level modules	All modules must be traceable.(ex. a single hierarchy which describes subordinate relationships between all modules)

4 Conclusion

A certification stands for not only a superiority of product, but prerequisite factor of sustainability. Most of COTS software companies are trying to get a certification endorsed by government agencies or professional institutions. For software industries, design recovery techniques could be used as the best approach which provide an effective and handy way for satisfying an compliance with standards such as ISO/IEC15408 and RTCA/DO-178B.

With the method provided in this paper, applicants for certification could deliver enough materials which are necessary to go through the process of certification in spite of the insufficiency or absence of data from software life cycle. The key processes of this method are digested as follows: first, planning and extracting a rough design structure with composition of external functions; second, defining sub-systems and classifying source codes; third, low-level analysis, and extracting low-level modules and interfaces; and fourth, high-level analysis with traceability. We expect our approach would be helpful to not only applicants who prepare a certification, but also evaluators willing to improve an evaluation process continuously with reverse engineering.

Limitation and further research issues are summarized as follows: 1) Case studies including actual data and evaluation processes should be provided to show various application cases; 2) Subdivision of each activities should be provided and described for practical use of the method; 3) Each activity should be revised continuously with accumulative feedback of use.

Acknowledgements

Acknowledgements from corresponding author, Minhyung Kim: I must express the most sincere gratitude to Dr. Kim for his unstinting helps and farsighted insight. Most of all, my deepest sense of obligation and gratitude goes to my parents who are everlasting shepherds.

References

1. TNO-ITSEF BV, Arrangement on the Recognition of Common Criteria Certificates in the Field of IT Security Papers (on the web: http://www.commoncriteriaportal.org/public/files/cc-recarrange.pdf). Netherlands Organization for Applied Scientific Research TNO (Accessed on April 2005)
2. ISO: Information Technology – Security techniques-Evaluation criteria for IT Security, ISO/IEC 15408. International Standard Organization (1999)
3. RTCA: Software Considerations in Airborne Systems and Equipment Certification, RTCA/DO-178. Radio Technical Commission for Aeronautics (1992)
4. Riverson, L. and Lingberg, B.: Reverse Engineering of Software Life Cycle Data in Certification Projects. DASC '03. The 22nd , Vol.1 (2003) 12-16
5. NIST, Common Criteria Evaluation and Validation Scheme (on the web: http://niap.nist.gov /cc-scheme). Information Technology Laboratory of NIST (Accessed on April 2005)

Information Privacy Engineering in ubiComp

Tae Joong Kim[1], In Ho Kim[2], and Sang Won Lee[3]

[1] Privacy Protection Team, Korea Information Security Agency
78, Garak-Dong, Songpa-Gu, Seoul 138-803, Korea
taej@kisa.or.kr

[2] Policy Planning Division, Korea Information Security Agency
78, Garak-Dong, Songpa-Gu, Seoul 138-803, Korea
kih@kisa.or.kr

[3] Dept. of Management Engineering, Korea Advanced Institute of Science and Technology
207-43, Chongyangni-Dong, Dongdaemoon-Gu, Seoul 130-012, Korea
sangwonlee@kgsm.kaist.ac.kr

Abstract. Ubiquitous computing demands a fundamental shift in the control of personal in-formation. Ubiquitous computing requires disclosure of personal information. The technology infrastructure means the basic components of the IT society such as web servers, DB, N/W, personal information devices, etc. And they can flow the personal information through the society. Business application systems which collect, store, process and/or disclose personal information should be identified and categorized. The process map for personal in-formation handling should document the flow of data from collection through all paths, internal and external, including all processing points, access/display points and destruction points. At this level, risks to personal information can be better understood. In this paper, we present the information privacy engineering for constructing to prevent the privacy invasion and measuring the economic value of privacy. We hope this privacy engineering will be used as one of the tool for protecting the users in ubiquitous age. The approach includes the followings; The Privacy Protection in Network, The database modeling in UbiComp, The privacy impact assessment in UbiComp.

1 Introduction

The social problem that personal information escapes has become powerful because security companies are professionally negligent of operating and managing personal information of their customers. In Korea, a company drained out personal information of their 4,000,000 or more customers. It is because the company was remiss in discharge of its duties in managing personal information. According to AOL/NCSA in the U.S.A. (November 2004), the 91 percent of Americans who use their personal computers at home have never heard about the spyware and most users of spyware never recognize it. [1] In this paper, we propose concrete criteria for technical or administrative security management in securing personal information.

2 Privacy Definitions

Privacy can have many aspects, but for purposes of this principle and the corresponding criteria, privacy is defined as the rights and obligations of individuals and entities

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 628–634, 2005.

with respect to the collection, use, disclosure, and retention of personal information. Personal information is defined as any information relating to an identified or identifiable individual. [2]

Firms must establish internal control and monitoring measures so that they may be enforced. For firms to prevent these risk potentials, it is desirable that they should put into three specific actions which would be difficult as the follows. The wireless world, in surfing the Internet or viewing e-mail, has the same privacy concerns as the wired world. The first consideration in developing a sustainable approach to privacy is the creation of an underlying enterprise data strategy. The second is the issue of technology infrastructures. The third element of a privacy strategy is business operations- the side of the triangle that shapes processes and procedures that guide consumer interactions. [3]

We present the privacy engineering in order to prevent the privacy invasion and measure the economic value of privacy. Now let us study on the followings; Tech Architecture, Legal Architecture, and Privacy Engineering.[4] The spyware is defined a program which stealthily intrudes a computer on internet, intercepts specific information, and tracks users' behaviors. And then, it stealthily transmits the information of users' activities to the third party. While a user setups a free software like freeware or shareware and one surfs specific websites, the spyware could read the user's information like emails. And when a spammer collects email addresses by use of a spyware or a user reads spam mails, some spywares could be installed.

Building the world's datasphere is a three-step process—one that we've been blindly following without considering its ramifications for the future of privacy. First, industrialized society creates new opportunities for data collection. Next, we dramatically increase the ease of automatically capturing information into a computer. The final step is to arrange this information into a large-scale database so it can be easily retrieved at a moment's notice. Once the day-to-day events of our lives are systematically captured in a machine-readable format, this information takes on a life of its own. It finds new uses. It becomes indispensable in business operations. And it often flows from computer to computer, from business to business, and between industry and government. If we don't step back and stop the collection and release of this data, we'll soon have a world in which every moment and every action is permanently "on the record."

3 Control Points

The personal information management systems protect personal information by use of intrusion detection systems and firewall systems. And the systems also provide additional backup and storage lest the access records should be forged or altered. It is necessary that the personal information (like password, biometric information, and so one) that authenticate users should be stored after encrypt it one-way (a decryption-preventive measure). It is sure that the encryption is an indispensable condition in transmitting or storing data. In personal information management systems and various personal computers of security companies, vaccine programs ought to be installed for scanning and clearing malignant programs such as computer viruses, spywares, and so on. In addition, the outputs of personal information should be minimized according to their usage and approved by chief privacy officer.

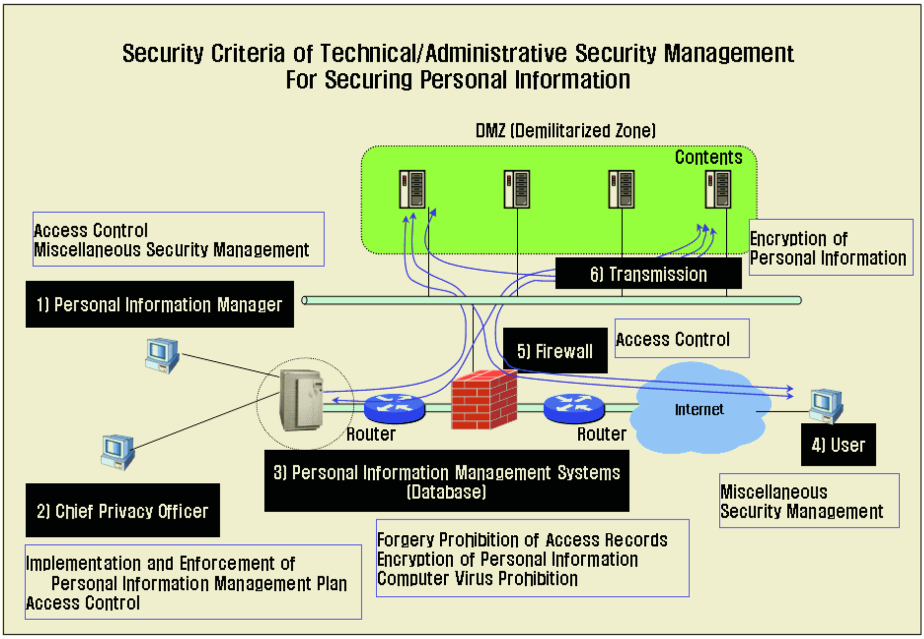

Security Criteria of Technical/Administrative Security Management
For Securing Personal Information
DMZ (Demilitarized Zone)
Contents
Access Control
Miscellaneous Security Management
Encryption of
Personal Information
6) Transmission
1) Personal Information Manager
5) Firewall
Access Control
Internet
Router
Router
4) User
3) Personal Information Management Systems
(Database)
2) Chief Privacy Officer
Miscellaneous
Security Management
Forgery Prohibition of Access Records
Encryption of Personal Information
Computer Virus Prohibition
Implementation and Enforcement of
Personal Information Management Plan
Access Control

Fig. 1. In technical or administrative security management for securing personal information, there are many security criteria. 1) Personal information manager: Access Control (Managing access control), Miscellaneous Security Management (Limiting the utilization of messenger or P2P, Prior approving when reproducing mobile media, Educating legal responsibility), 2) Chief Privacy Officer: Implementation and Enforcement of Personal Information Management Plan (Organizing and administrating the department of personal information security, Managing users of personal information, Taking the technical measures for blocking illegal access), Access Control (Setting up regulations for password and putting them into operations, Maintaining updated access accounts), 3) Personal Information Management Systems: Forgery Prohibition of Access Records (Setting up applications for encryption/backup/integrity, Storing/verifying /supervising access records), Encryption of Personal Information (Encrypting authenticated information), Computer Virus Prohibition (Setting up vaccine softwares, Updating legacy softwares), 4) User: Miscellaneous Security Management (Minimizing outputs, Limiting setting up information collectors), 5) Firewall: Access Control (Setting up intrusion detection systems and firewall systems), 6) Transmission: Encryption of Personal Information (Encrypting and transmitting major personal information)

3.1 Control Point in Browser

The spyware program announces about service conditions or personal information supply while it is being installed usually. The announcement, however, is apt to guide most users to let them ignore the contents of it and click the consent without any notices. In using internet, users should accustom them to downloading and executing only the program that its function and purpose are precise. Users tend to select “yes” or “no” in answering the questions like “Do you want to setup and execute ***?” while they encounter some programs in surfing on the internet. They don’t know why they need the programs, even though the 50 percents of programs are spywares. In using their computer, users must be fully aware that the yes is the agreement on the question, “Do you want to setup and execute ***?”

3.2 Control Point in Network

The access control cuts the unauthorized accesses off from the personal information management systems. And the unauthorized accesses stand for the illegal usage, leakage, and forgery of the personal information. In most cases, the system users are classified and approved according to the security policy such identification and authentication. The identification and authentication is a most basic means for access control to secure system resources. The responsibility for personal information security should be maintained from employment to retirement of human resources. And then the most basic security policies should be also prepared for protecting and confronting many risks like the leakage or forgery of information intentionally and accidentally.

3.3 Control Point in Database

Proper authentication is a critical component of DB modeling. It is because if once a party has been accepted into the system, a legally binding transaction process has begun. Database Security Control consists of the Flow Control, the Inference Control, and the Access Control. DB Security Control measures restrict who gets in, manage their identities and access rights, Column level controls, and Audit to hold users accountable for their actions. Classification of Security Threats according to the way they occur: (1) Accidental; Human errors (incorrect input, incorrect use of applications), Errors in software (incorrect application of security policies, denial of access to authorized users), Natural or accidental disasters (damage of hard-/software) (2) Intentional; Authorized users who abuse their privileges and authority, Hostile agents (improper users, insiders / outsiders) executing improper reading or writing of data legal use of applications can mask fraudulent purpose. (Viruses, Trojan Horses, Trapdoors) [5]

4 Security Implementation

A full assessment identifies both gaps in the design of the privacy program in considering business risks and opportunities, as well as gaps in the operating effectiveness of current policies and practices. It tells you what will never work because of flawed design and what isn’t working because of flawed execution. The resulting strategy and plan will document recommended solutions as well as how to ensure those solutions are implemented. [6]

In the final analysis, the assessment should set forth a thoughtful strategy and plan for closing the gaps and achieving the goals and objectives. Findings and recommendations should be presented in consideration of the organization’s culture and values as well as risk management priorities. Finally, consideration must be given to overcoming anticipated barriers to change, as well as the trade-off between cost and benefit. Now, with plan in place, it is time to implement. [6]

The implementation method of related laws should be linked to Privacy Engineering. In designing database, the constraints for securing privacy should be defined at the early stage.

4.1 Security Implementation in Browser

Users who have no expertise can cope with spywares by use of anti-softwares which detect spywares. In this case, users are under an obligation to purchase anti-spyware products and use them periodically because spywares intrude into users' personal computers continuously. Users must pay steady attention to clear spywares lest they should mistake proper programs for spywares and cleaned spywares should be restored. It is necessary to install a program when it is necessary and not to download a program that is unnecessary. The additional setup that limits disapproved programs is a compulsory action. The spywares intrude into users' systems before one knows. Users' should not click unnecessary popup-links and should delete all the spam mails whenever detected.

4.2 Security Implementation in Network

Personal Information Management Systems administrate Intrusion Detection Systems and Firewall Systems for strengthening network security so as to protect them from illegal access activities like hacking. Hence, the systems protect themselves against various intrusions such as worm, virus, hacking that have a bad effect on information and communication network. It is an aid to build up the security of major electrical circuits connected to external networks. The transmission should get everything in readiness for encrypting personal information (like social security number, password) and preclude it, which is supported by network encryption methods such as Secure Socket Layer, Secure Electronic Transaction, and so on. A company stores the approval information (password, biometric information …) after encrypting it one-way not to be decrypted.

4.3 Security Implementation in Database

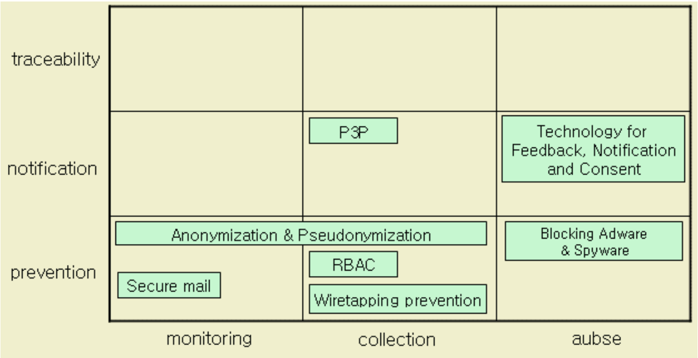
To ensure successful identity management, a digital identity solution should support at least the following basic requirements. (1) Reliability and dependability: Identity theft is one the fastest growing electronic crime and it is expected to accelerate. Digital identity must offer protection against forgery and related attacks. (2) Controlled information disclosure: Users must be given control on what identity to use in specific circumstances. Control must also be given with respect to possible replication and misuses of the identity information a party reveals in a transaction. (3) Mobility support: The mobile computing infrastructure can keep track of an individual's physical location. In addition, mobile computing bears some peculiarity such as limited bandwidth and limited display size. [7]

There are several ways to look at approaches to identity management. One may look at this question as a matter of policy and law; or as business cases and practices;

or as technical architectures and technologies; or even at guiding philosophies and principles. [8] We must pay attention to understand how business process controls and identity management work together, which could ensure that the identity management and its supporting infrastructure are delivered within the context of business objectives.

5 Conclusions

Specific risks of having inadequate privacy policies and procedures include; Damage to the organization’s reputation, brand, or business relationships, Legal liability and industry or regulatory sanctions, Customer or employee distrust, and Disruption of international business operations. The future is uncertain. Firms can use scenario planning so as to develop privacy strategies in the technological, regulatory, and competitive landscape.[4]



	monitoring	collection	aubse
traceability			
notification		P3P	Technology for Feedback, Notification and Consent
prevention	Anonymization & Pseudonymization Secure mail	RBAC Wiretapping prevention	Blocking Adware & Spyware

Fig. 2. In order to strengthen the protection of personal information, a foundation of laws and polices must have been established. Also, the government should prepare the related technology which can handle the privacy issue. In this process, we have to make an effort not to omit the whole technology to handle the increasing the future privacy issues. We suggest the following technological framework to prepare the safe ubiquitous computing environment. [9]

However as we can see from the above framework, we don’t have proper technology with traceability issue. As our society advances to ubiquitous computing environments, more privacy related arguments can be occurred in the future. If we encounter such problem, we have to decide who’s is guilty and who’s innocent through technologies. These technologies would have a function of traceability and auditability (e.g. forensics in these days). For this reason, we believe that the future computing privacy technologies need to be focused on developing the solution for this area.

References

1. AOL/NCSA : Online Safety Study(2004)
2. US: US Safe Harbor Privacy Principles (2000)
3. Deloitte Research: Creating a privacy value strategy (2002)

4. TJ Kim, SW Lee, EY Lee : Privacy Engineering in ubiComp(2005)
5. Addison-Wesley: Database Security (1994)
6. Glasser Legalworks: Privacy- Taking Action to Safeguard Customer Loyalty (2001)
7. Information Society Technologies: Identity management PIM Roadmap- Multiple and Dependable Identity Management- R & D Issues (2002)
8. The National Electronic Commerce Coordinating Council: Identity Management- A White Paper (2002)
9. TJ Kim, EY Lee, IH Kim, KI Yoon, YJ Kim: Architecture of the Privacy Governance in ubiComp (2004)

Design and Implementation of Home Media Server for Personalized Broadcasting Service in Ubiquitous Environment

Chang-ho Hong[1], Jong-tae Lim[2], Chang Sohn[1], and Ha-eun Nam[1]

[1] Digital Research Center, DAEWOO Electronics Corp. ASPD B/D.
254-8 Gongdok-dong, Mapo-gu, Seoul, Korea
{chhong,sohn,namhe}@dwe.co.kr

[2] School of Electronics, Telecommunication and Computer Engineering
HANKUK Aviation University
lim@hau.ac.kr

Abstract. Home Media Server(HMS) as implemented can retrieve metadata of multimedia contents using TV-Anytime in ubiquitous environment. HMS has various functions using user preference like genre, publisher, keyword and usage history of content is consumed. We introduce the design and implementation of Home Media Server for ubiquitous environment using simple object access protocol and universal description discover & integration service.

1 Introduction

A expansion of digital broadcasting enabled personalized broadcasting and there are also increasing demands for personalized broadcasting. Particularly with the popularization of the Internet, users became able to access broadcasting service servers through return-channel and, as a result, people are getting more interested in bi-directional personalized broadcasting service.

Users can obtain detailed information about programs such as title, synopsis, schedule and review, and search specific programs using metadata received in unidirectional broadcasting environment. Metadata expressing such information were standardized in TV-Anytime Forum [1]. In unidirectional broadcasting environment like PSIP[6], however, the volume of metadata transmitted to users is limited due to limited bandwidth allocated to the transmission of metadata and, accordingly, services provided to individual users are also restricted. Moreover, for the provision of services well-customized to individual demands, it is necessary to deliver information and requests of individual users to personalized service providers. If, for the delivery, a communication function is added to broadcasting HMS through return-channel, bi-directional personalized service can be provided. Representative examples of bi-directional personalized service are a person's retrieval of personalized programs based on the individual user's preference or choice, recommendation of personalized programs based on individual users' history of service uses. The standard for metadata communication between user HMS and metadata servers for bi-directional personalized service was also established in TV-Anytime Forum [2].

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 635–641, 2005.

As shown in Figure 1, the general structure of bi-directional personalized service including the implemented bi-directional personalized broadcasting HMS is composed of digital terrestrial publisher, metadata server, broadcasting contents server, bi-directional personalized broadcasting HMS and UDDI[3] service registry server. The metadata server manages various types of information about contents, and the UDDI registry server manages different kinds of services provided by metadata servers and information about how to access the services. A metadata service provider, who is separated from digital broadcasters, may run an independent metadata server or a digital broadcaster can run a metadata server directly. Bi-directional personalized broadcasting HMS receives and records digital programs broadcasted by digital terrestrial publishers and, in response to users' requests for personalized broadcasting service, it accesses the metadata server and exchanges metadata necessary for bi-directional personalized service.

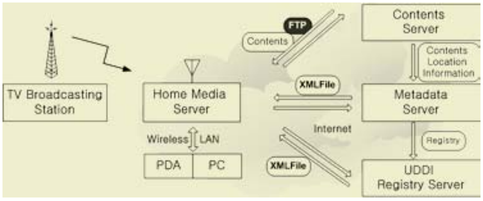


Fig. 1. The overall structure for bi-directional broadcasting service

2 Metadata Service in Bi-directional Broadcasting Environment

Metadata can be transmitted through unidirectional broadcasting or a bi-directional network. The transmission of metadata through a bi-directional network is more advantageous in several points than that through unidirectional broadcasting. For example, metadata can be provided according to users' personal demands. A larger volume of metadata can be transmitted in a bi-directional network than in unidirectional broadcasting, which has a limitation in bandwidth. Moreover, users can be provided with metadata without receiving broadcasting programs. Bi-directional metadata service means the exchange of metadata between broadcasting HMS and metadata servers using the return-channel and it is divided into metadata retrieval and user-centric metadata transmission [2]. Metadata transport protocols are SOAP [4] and HTTP [5]

2.1 Metadata Retrieval and Transmission of User-Centric Metadata

As shown in Figure 2 (left), metadata retrieval means a user's request for specific metadata service through bi-directional personalized broadcasting HMS and, in response to the request, the provision of metadata by the metadata server. For example, in order to retrieve specific programs, a user requests the metadata server to provide information about CRID (Content Reference Identifier), title, genre, keyword and publisher of contents. The request is made through the operation of TV-Anytime called 'get_Data' and the user is provided by the server with information.

User-centric metadata is transmitted through the operation of TV-Anytime 'submit_Data.' Metadata related to the user such as UsageHistory and UserPreferenc is sent to the metadata server. On receiving the data, the metadata server sends back 'Acknowledgement' for the user-centric metadata as in Figure 2 (left). UserPreference is metadata about the user's personal preference, and UsageHistory about the user's usage history. Using UsageHistory metadata, which show how the user has consumed contents, the metadata server understand the user's pattern of consumption and, based on the understanding, recommend suitable contents or provides targeting services. Figure 2 (right) shows the general structure of UsageHistory information.

In order to create UsageHistory metadata, broadcasting HMS must update the user's UsageHistory information such as contents watched by the user and time when the user watched the programs. More efficient UsageHistory metadata may include the user's methods of consuming contents, e.g. the operation of Play, Preset Recording, Pause and FF/REW, the time of operation and the frequency of function.

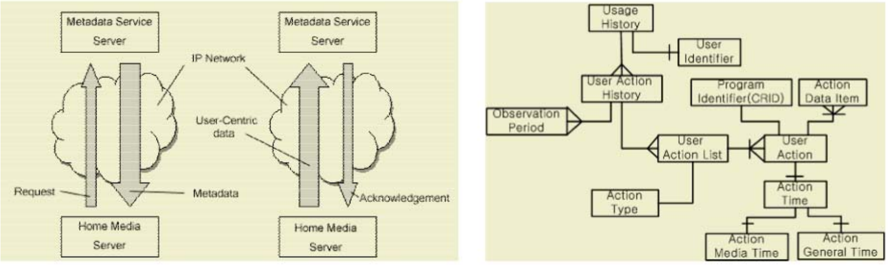


Fig. 2. Bi-directional Metadata Services(left) and UsageHistory Data Structure(right)

2.2 Universal Description Discovery and Integration

Network module uses a transport stack explained in Section 3 of Chapter 2 in order to exchange metadata through bi-directional communication. In addition, it receives transport streams through FTP, and delivers the state of data transmission and errors.

Figure 3 shows the process of metadata service retrieval. Bi-directional personalized broadcasting HMS sends a query to the UDDI service registry server using <find_business>, and receives metadata containing <businessKey> and <serviceKey> from the UDDI service registry server. Broadcasting HMS parses the metadata, extracts <serviceKey> of TV-Anytime from the data, and sends a query to the server in the form of <get_ serviceDetail> in order to get an access point to the metadata server for the service. Then, from the replied document, broadcasting HMS extracts the IP of the metadata server, which is the access point of metadata service, and connects to the metadata server. If connection is accepted, HMS sends the server a XML [7] document corresponding to 'get_Data' to retrieve information from the server. Then the metadata server sends a reply to the request to the personalized broadcasting HMS.

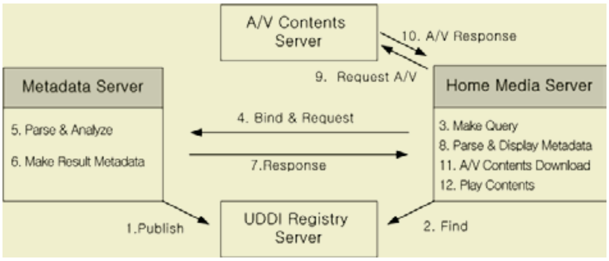


Fig. 3. Metadata retrieve and Universal description discovery & integration service

3 Implementation of Bi-directional Personalized Broadcasting HMS and an Example of Service

Bi-directional personalized broadcasting HMS can receive and record digital terrestrial broadcasts, and provides various bi-directional personalized service through network module. If the user wants to obtain information about programs selected from contents that are currently being broadcasted, the system receives data on the programs from the user and requests information to the metadata server. It again receives a broadcasting list and detailed information meeting the user's request and provides them to the user so that the user can select and watch preferred broadcasts. In addition, the system provides bi-directional personalized services by sending user-centric metadata UserPreference or UsageHistory to the server and receiving the server's recommendation of programs fit for the user.

Bi-directional personalized services implemented in the present broadcasting HMS are explained below. We explains the service of sending a user's preference to the metadata server and receiving from the server recommendation of programs fit for the user's preference, the service of sending a user's UsageHistory from database and receiving from the server recommendation of programs based on the user's tendency, and an example of metadata service retrieval.

3.1 Broadcasting Program Retrieval Using User Preference

Metadata on UserPreference contains preferred genres, program schedule, publishers, keywords. These items are directly entered by the user when the user account is opened in broadcasting HMS, and may be changed when the user retrieves programs. Figure 4 (left) shows a screen for entering preference data. In the figure, three preferred genres Fiction, Sports and Leisure were selected. Users who want to select keywords or publishers can use the Detail Information ('More') menu on the right bottom.

Based on preference stored or entered as above, a program retrieval request is prepared as a XML document and sent to the metadata server. Then, in reply to the request, the server sends a reply XML document to HMS. Figure 4 (right) shows the server's reply to the preference. It shows a list of programs of the preferred genres, namely, fiction, sports and leisure, broadcasting time, broadcasters, download ability.

In the figure, recommended programs are shown together in a list. If the Recommended Genre menu on the left side of the screen is selected, recommended programs only for the selected genre can be viewed. Download ability tells whether or not the contents can be downloaded from the contents server at <ProgramURL> as specified in ProgramLocationTable of transmitted metadata. In addition, if the user wants detailed information about each of recommended programs, the HMS may request the information to the metadata server. Figure 5 (left) shows detailed information about program 'Travel Show! Escape from Routine' such as its synopsis and review sent by the server in reply to a request for detailed information. As shown in Figure 5 (right), a list of recommended programs personalized to each user's consumption pattern is provided to the user together with information such as program title, publisher, broadcasting time, length and download ability, and the user can select a program from the list and watch it or do preset recording

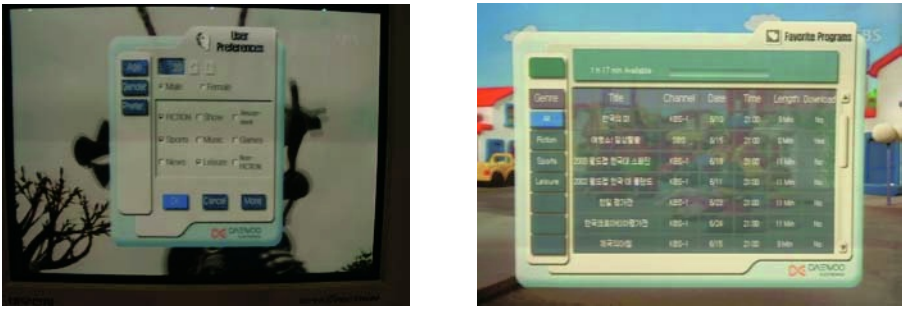


Fig. 4. Snapshot of setting the user preference (left) and Snapshot of metadata service server response for user preference (right)

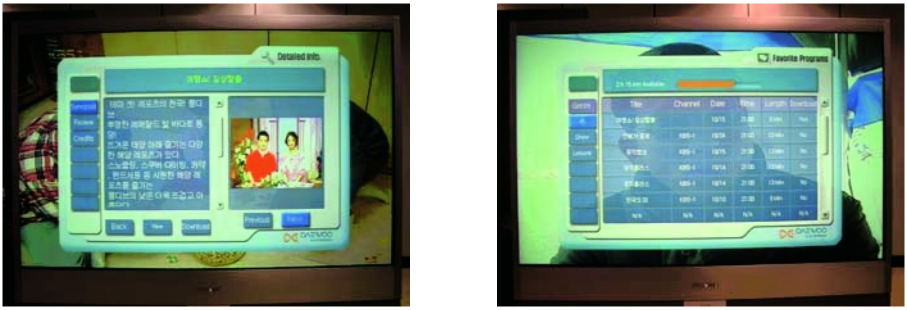


Fig. 5. Detailed Information of selected program (left) and snapshot of program list recommended by metadata service server (right)

3.2 TV-Anytime Metadata Service Retrieval

The system sometimes needs TV-Anytime metadata service but does not have information about metadata servers. In this case, it can obtain information about metadata servers through metadata service retrieval using UDDI. If <find_business> query document is sent to the UDDI service registry server, the UDDI server sends the corresponding provider's <businessinfos> and <serviceinfos>. HMS parses the XML document using libXML [8] from the UDDI server, extracts <serviceKey> included in <serviceinfos>, creates <get_serviceDetail> query document as shown in Table 1, and sends it to the UDDI service registry server. Contents in the rectangular box show <serviceKey> information. The server's reply to the query is presented in Table 2. The XML document is parsed and URL (Uniform Resource Locator) for the metadata server is obtained. URL for the metadata server in the second rectangular box in Table 4 is http://211.236.101.57/TVAService. Then, using the URL, personalized broadcasting HMS connects to the metadata server, and exchanges metadata for bi-directional personalized service.

Table 1. Example of <get_serviceDetail> query

```
    <?xml version="1.0" encoding="UTF-8"?>
    <soap:Envelope xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/">
      <soap:Body>
        <get_serviceDetail generic="1.0" xmlns="urn:uddi-org:api">

         <serviceKey>8e9cd030-5ca7-4c12-a79a-4074f2905894</serviceKey>

          </get_serviceDetail>
      </soap:Body>
</soap:Envelope>
```

Table 2. Example of received XML document include <accessPoint>

```
    <?xml version="1.0" encoding="utf-8"?>
    <soap:Envelope xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema">
    <soap:Body>
      <serviceDetail generic="1.0" operator="Microsoft Corporation" truncated="false" xmlns="urn:uddi-org:api">
       <businessService serviceKey="8e9cd030-5ca7-4c12-a79a-4074f2905894" businessKey="76dcbd4c-c879-4938-aa05-
    3787e05cc939">
      <name>tv-anytime-org:get_Data_v10</name>
             <description xml:lang="en">TV-Anytime WSDL interface for get_Data port</description>
         <bindingTemplates>
            <bindingTemplate bindingKey="791e23bf-dd2e-45ec-a8fb-4a98a01b6185" serviceKey="8e9cd030-5ca7-4c12-a79a-
4074f2905894">
              <description xml:lang="ko">TVAny time metadata service by softchaser</description>

         <accessPoint URLType="http">http://211.236.101.57/TVAService</accessPoint>

       <tModelInstanceDetails>
                <tModelInstanceInfo tModelKey="uuid:9c8976d2-9ef4-498c-88fe-6fdad9c28f02"/>
                  </tModelInstanceDetails>
             </bindingTemplate>
          </bindingTemplates>
       <categoryBag>
          <keyedReference tModelKey="uuid:c1acf26d-9672-4404-9d70-39b756e62ab4" keyName="Specification for a web service
described in WSDL" keyValue="wsdlSpec"/>
          <keyedReference tModelKey="uuid:c1acf26d-9672-4404-9d70-39b756e62ab4" keyName="Specification for a web service
using SOAP messages" keyValue="soapSpec"/>
        </categoryBag>
        </businessService>
      </serviceDetail>
     </soap:Body>
    </soap:Envelope>
```

4 Conclusion

In order to utilize bi-directional personalized broadcasting service, the present study presented an example of bi-directional personalized service by designing and implementing bi-directional personalized broadcasting HMS that satisfies TV-Anytime standard. The implemented bi-directional personalized broadcasting HMS sends metadata about individual users' choice or preference and their UsageHistory to metadata servers through the return-channel, receives replies from the servers, and provides personalized services to the users based on the information from the metadata servers.

Bi-directional environment enables services personalized to each individual's demand, which are not available in unidirectional environment. The activation of bi-directional personalized services must be preceded by the expansion of TV-Anytime metadata service providers who provide services diversified and specialized for different groups of users.

References

1. SP003v1.3, TV-Anytime Specification-Metadata, 2003, "http://www. tv-anytime.org"
2. SP006v1.0, TV-Anytime Specification - Metadata Services over a Bi-directional Network, 2003, "http://www.tv-anytime.org"
3. Universal Description Discovery & Integration, Version 3.0, "http://uddi.org/pugs/uddi-v3.00-published-20020719.htm"
4. Simple Object Access Protocol (SOAP) 1.1, W3C Note, 8 May 2001, "http://www.w3.org/TR/2000/NOTE-SOAP-20000508"
5. RFC1945- Hypertext Transfer Protocol, HTTP/1.0, "http:// www.ietf.org/rfc/rfc1945.txt"
6. ATSC Standard A/65. Program and System Information Protocol for Terrestrial Broadcast and Cable, http://www.atsc.org"
7. W3C, Extensible Markup Language (XML) Version 1.0 Recommendation, February 1998, "http://www.w3.org/TR/2004/ REC-xml-20040204"
8. "http://www.xmlsoft.org"
9. "http://www.mysql.com"
10. Lee.j.s. Lee.suk.phil "Design and Implementation of bi-directional TV-Anytime system for personalized broadcasting service" The Korean Internet Broadcasting / TV Institute Vol 3, No.1, 2003.
11. Z. Yu and X. Zhou, "TV3P: An Adaptive Assistant for Personalized TV," IEEE transaction on Consumer Electronics, Vol. 50, No. 1, Feb. 2004

A Distributed Approach to Musical Composition

Michael O. Jewell, Lee Middleton, Mark S. Nixon,
Adam Prügel-Bennett, and Sylvia C. Wong

University of Southampton, Southampton, SO17 1BJ, UK
moj@ecs.soton.ac.uk

Abstract. Current techniques for automated composition use a single algorithm, focusing on one aspect of musical generation. In our system we make use of several algorithms, distributed using an agent oriented middleware, with each specialising on a separate aspect of composition. This paper describes the architecture and algorithms behind this system, with a focus on the agent framework used for implementation. We show early results which encourage a future application of this framework in automated music composition and analysis.

1 Introduction

Traditionally, algorithmic music composition has aimed to create independent pieces of music using rule-based techniques. These techniques do not have to be handled by a computer - they have been used as far back as 1026 when Guido d'Arezzo assigned pitches to vowel sounds [1] and, more recently, when Ron Pellegrino created music using light hitting wall-mounted photoresistors. However, as computers are now sufficiently powerful, algorithms are often carried out in software.

The burst of computational composing algorithms began with Arnold Schönberg at the start of the 20th century, with Webern and his successors forming serialism from these roots. Iannis Xenakis was a pioneer who, from his 'succès du scandale' *Metastaseis* in 1955, produced multimedia creations based on probability, sonic phenomena, texture, and random generation, and this work contributed to the stochastic approach of composition [2]. Further approaches, such as Voss and Clarke's fractal techniques [3], McAlpine's cellular automata method [1], and Burton's genetic algorithm systems [4] followed, and these make up a collection of 'stock' composition methodologies.

Until recently, however, music composition has concentrated on using single algorithms to control individual elements of the generation process. For example, an algorithm is used to create chord patterns, while the melody, rhythm, and keys are set a priori. Our new distributed approach aims to treat the existing algorithmic techniques as building blocks for the creation of a music composition system, where different algorithms can be plugged in for evaluation.

The distributed composing framework is further bolstered by its strong ties to other media. Rather than generating music with no prior information, a composer model[5] is used to provide a priori information for the algorithms, and a script representation allows for the alteration of these parameters at pertinent points in the bound medium.

This paper is split into two halves. The first provides a technical background to the Light Agent Framework, which is at the core of our system, while the second describes

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 642–648, 2005.

how this framework is used for distributed music composition, with some preliminary results given. Finally, we suggest how we will take advantage of the system for future agents.

2 The Light Agent Framework

The main intention behind our agent framework was simple - that it should be lightweight. To maximise uptake it had to be intuitive for a user to package algorithms, both existing and original, and furthermore it had to be undemanding on the host computer, hence providing the maximum possible resources to the agent.

However, while a streamlined approach was appealing, it was decided to allow for the possibility of adding extra features without disturbing the original interface. This is especially true of the router, which is covered in more detail later in this section. Balancing features and simplicity was key to the development of our system, and we believe it is therefore suitable for a wide variety of applications.

2.1 Agent Design

At the core of every agent in the LAF system is an engine. This was initially designed as a separate module, but was later subsumed into the agent class, partly for ease of threading, and partly to reduce the files required to design an agent. The engine can only be activated by one client at a time (to preserve atomicity) but this is complicated when several parameters are passed at dissimilar times. As such two messages, LOCK and UNLOCK, were created.

When any agent connects to a router, it sends a ‘stub’ of information. This includes the agent type, the creator name, a description, a version number, and any inputs or outputs that can be accommodated. If a client then wishes to lock an agent, it requests that the router select an unlocked agent, return a unique name, and then prevent other clients from interacting while the agent is locked. Once the client is finished, or if it disconnects unexpectedly, the router can unlock the locked agent to let other clients make use of it. Most importantly, when an agent is locked, only the locking client can alter the input and output parameters. To all other clients, the agent appears immutable.

Ports. The modeling of the agent parameters uses a further design feature of the light agent framework - ports. Ports have unique names, with a loose hierarchy provided by ‘.’s. Five port subcategories are defined:

1. agent.identity.*
   The identity ports are a structured representation of the stub described in the previous section, with name, type, creator, description, and version fields. The name is not set by the user; instead the router generates a unique name from the type when the agent connects. As with port names, type names are hierarchical in nature, with dots delimiting. For example, ‘string.concat’ is a valid type, as is ‘music.composing.genetic’.

2. agent.input.* and agent.output.*
   The input and output ports are responsible for passing data between agents and clients. The input ports are immutable, whereas the output ports can be altered to indicate the results of a process. Agents can be configured to require certain ports, while others can be set as optional. This further allows for the chaining of agents, with execution triggering when all compulsory ports are initialized.
3. agent.state
   The agent.state port is the simplest in the framework, and contains a string representing the state of the agent. This can have one of four values, namely 'waiting', 'ready', 'running', and 'exiting'. The agent enters the waiting state on connection, the ready state when all compulsory ports are set, the running once executed, and the exiting state when disconnected. If monitored for changes, clients can use this to give a high level status indication.
4. agent.call.*
   When running larger tasks, it is useful for clients to be able to monitor the progress of jobs. The call ports facilitate this and provide a lower level alternative to the state ports. Four ports are provided: percent, time.current, time.total, and status. agent.call.percent provides the percent of the job complete as a double value, agent.time.current and agent.time.total provide the duration of the current job and the time that the agent has been locked, while agent.call.status is a string describing the state of the agent. The percent and time.* ports are most likely to get updated often, so these are typically polled at set intervals. The status, however, is more suitable for an interrupt approach.

Ports are typed, but these types can be defined by the agent author. The standard base types are provided - Boolean, integer, string, double - and these include validation functions to ensure that no incorrect parameters get passed through to the engine.

Remote Agents. To ease the usage of agents by clients, a remote agent interface is provided. This acts as a proxy, and allows for methods to be called on an agent as if the instance was local. All communications are handled via an agent session, connected to the router, and when an agent is locked a remote agent is returned to the client. Only a select few methods are provided, including call (to commence executing the agent), port set and retrieval methods, and functions to request information on the locked agent, such as its port and IP address. Once a remote agent is finished with, it can be unlocked via an agent session method, thus eliminating the need to communicate with the router directly.

2.2 Router Design

As mentioned previously, the LAF router is modular in design. Several plugins were implemented for the Java router, including a logging plugin, a monitor plugin, an identification plugin, and a state plugin. The first three of these correspond to the LOGGER/LOG, MONITOR/NOTIFY, and IDENTITY/IDENTIFY messages. Respectively, these message pairs allow for the transmission of logging information, notification information on a port change, and agent stub details. The abstraction of these messages into removable components allows for a very lightweight router for circumstances where resources are limited. Finally, the state plugin is responsible for keeping an accurate representation of the state of the router, such as which agents are connected and the states of these agents. This is primarily for debugging and audit trails, but can also be useful for web-based status monitoring.

Further to these plugins, the router implementation uses a 'selector' module. This specifies which agent should be selected when a client requests a type. The base model in LAF is that of the locking selector. This handles the LOCK/UNLOCK messages, and locks the next available unlocked agent in order of their subscription. This could be extended to allow for resource or platform checks. The latter case is especially suited for the launching of agents on machines with sufficient resources.

In summary, the basic router only handles subscription messages, disconnect messages, and routing itself. It is through the use of plugins and selectors that features can be added and as such the router can be tailored to suit the application.

3 Agent-Based Composition

To create our composing system, the process of composition was split into individual tasks. This keeps the system analogous to the traditional approach for music writing. Each of these tasks was then implemented as an agent, allowing for the production of an agent graph connecting them together. Furthermore, this gave the ability to rearrange the system to test different combinations of agents. This section details the standard features in our composition agents, then focuses on the operation of these agents.

3.1 Agent Structure

The agents in our composition system all follow the same model. Each is in the 'music.composing.' hierarchy to distinguish from other agents, and each has 2 inputs and 2 outputs. The landmark port and the MusicXML [6] port are the two inputs to the agents, and modified versions of these arguments are provided on the output.

The landmark file contains sets of meshes and mappings, placed at key points in the source medium. In film, a key point may be where the location changes or a character's personality alters. This information is used to prime the algorithms within the individual agents, with the output port containing modified parameters if necessary (for example, to include the beat information). The file is split into segments, to allow for scenes and shots, and these sections can be defined in frames or seconds.

Where the landmark is used for input, the MusicXML port is the output. MusicXML is used as the music transfer format, as it is both easily parsed and able to contain a high level of detail. Where MIDI represents only 'note on' and 'note off', MusicXML contains structures to describe a wide range of musical attributes, such as note length and articulation. We store the music in a part-wise approach, as each agent can then work on individual parts.

To handle the two inputs, each agent has an initial parsing step. This parses the landmark and MusicXML values into component objects. Once execution has completed there is a final parsing step, where the landmark and MusicXML structures are altered to include the results of the operation, and then they are serialised and the output ports are set. This standardised structure, as shown in Figure 1, simplifies the interconnecting of several agents via the agent graph system, and eases debugging - as only two parsers are required.

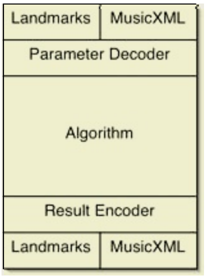


Fig. 1. The standard structure used to represent a musical agent

3.2 Agent Implementation

At present, seven agents have been implemented in the SBS system - namely tempo, pulse, instrumentation, key, chord, rhythm, and melody. Of these seven, six use genetic algorithms to produce the final results, while the other (tempo) only uses a genetic approach when no tempo is provided or when a tempo cannot be easily calculated from event information. These agents, as well as the remaining agents that are under development, are connected to one another in the arrangement shown in Figure 3 using the agent graph system.

The landmark file, described earlier, provides the means for the agent fitness functions to evaluate the population during composition. The musical element is stored in a chromosomal representation; for example, the melody agent stores pitch values within each gene. The tempo, pulse, instrumentation, and rhythm agents use a histogram-based parameter which specifies the probability of each alternative occurring. Taking instrumentation as an example, stringed instruments may have a high probability of occurring, whereas trumpets might have a low probability. In these cases it is trivial for the genetic algorithm to evolve the population in such a way that the most likely combination of options is produced.

The other cases (namely key, chord, and melody) are more complex, as these use a Markov model approach to fitness evaluation. The Markov model specifies the probability of moving from one state to another, with probabilities defined by the composer representation. This state-based approach is essential to these agents, with key changes, chord progressions, and scales being central to the creation of satisfactory music. Again, however, it is not difficult for the genetic algorithm to employ these parameters as constraints in the fitness function. The fitness of the chromosome can be calculated by stepping through the genes, totalling the probabilities of moving from one state to another, with higher results suggesting that the chromosome is more suitable.

The design of the SBS system is such that each stage adds detail to the prior stage: for example, the tempo agent determines the location of beats within the music (and hence the speed), which the pulse agent then supplements with strength information to indicate barline placements. Figure 2 shows the results of this process, with a generated pulse augmented by rhythm and melody.

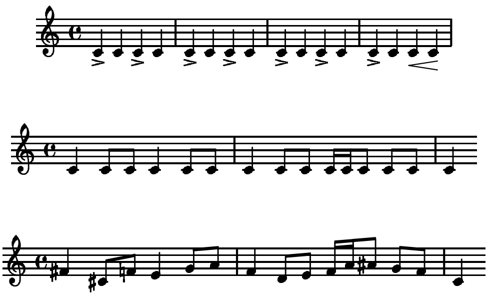


Fig. 2. The creation of a musical phrase using three connected agents. First a pulse is generated, which is then used to generate a suitable rhythm. Finally, a melody is generated and overlaid on the rhythm

The majority of the stages can also be parallelized to increase efficiency: a film can be split into scenes, and each scene can be handled by a separate collection of agents. There is a need for consistency however, so future work will include sharing musical themes between agents. The Light Agent Framework is ideal for this implementation, as a single router can ensure that messages are passed efficiently between the various agents, as well as selecting available agents to process the landmark representations.

4 Conclusions

The Light Agent Framework is a very elegant approach to a difficult problem - the distribution of applications using multiple components. While typical systems can require extensive configuration, our framework allows for applications to be coded away from the API and then simply dropped in as agents. The framework is also very portable, with versions available in Java, C++, and Python, all using the same user-side function prototypes.

When applied to music composition, LAF is the ideal solution to allow for the task. Though current approaches to music composition are largely sequential the generation of music is by its nature distributed, and our implementation of genetic algorithms encapsulated within individual composing agents proves this fact. Our system moves away from the typical monolithic approach to automated composition, and also allows for the direction of the music by existing temporal media. Furthermore, the framework provides the facilities necessary for the parallelisation of many of the composition tasks, hence reducing processing time.

Two further agents, one for music validation and one for music rendering, are already at the design stage, so we hope to obtain complete generated scores for further testing. The validation stage is particularly suitable for the agent framework, as we plan

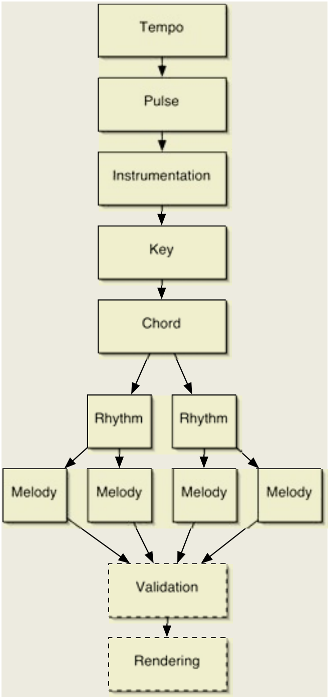


Fig. 3. A structure for the distributed composition of music. Future agents have dashed borders

to generate a large number of possible scores and use these as the bootstrap for a genetic algorithm. This will require tens, if not hundreds, of melodies, and hence will involve many melody agents communicating with several accumulation agents, with these communicating with the validation agent.

References

1. McAlpine, K., Miranda, E., Hoggar, S.: Making music with algorithms: A case-study system. Computer Music Journal **23** (1999) 19–30
2. Harley, J.: The electroacoustic music of Iannis Xenakis. Computer Music Journal **26** (2004) 33–57
3. Voss, R.F., Clarke, J.: 1/f noise in music and speech. Nature **258** (1975) 23–33
4. Burton, A.R., Vladimirova, T.: Generation of musical sequences with genetic techniques. Computer Music Journal **23** (1999) 59–73
5. Jewell, M.O., Nixon, M.S., Prügel-Bennett, A.: CBS: A concept-based sequencer for soundtrack composition. In: WEDELMUSIC. (2003)
6. Good, M.: MusicXML: An internet-friendly format for sheet music. In: XML Conference and Expo. (2001)

Efficient Mobility Management Using Dynamic Location Register in IMT-2000 Networks

Il-Sun Hwang, Gi Sung Yoo, and Jin Wook Chung

Dept. of Electrical and Computer Eng. SungKyunkwan Univ., Korea
his@kisti.re.kr

Abstract. In this paper we propose efficient mobility management using Dynamic Location Register (DLR) method, which is efficient for smaller cell and more frequent terminal moving in boundary cell in IMT-2000 networks. Whenever a terminal moves to another Registration Area (RA), Location Register (LR) circle change dynamically, and the terminal can track location by querying DLR of current terminal when a call originate. The proposed scheme reduces location traffic compared with IS-95 standard.

1 Introduction

The IS-95 and GSM [2,3,5] based mobility management scheme which records all the movements of users in a centralized DB, Home Location Register (HLR), is questionable considering that keeping track of lots of users in real time is not a simple task. This scheme has been the bottleneck problem on HLR which occurred in due lots of signal transfer between one HLR and many Visiting Location Registers (VLRs) and Ping-pong effect which arise frequently in the boundary of RA because of terminal's Zig-Zag movement or Ping-Pong effect. For this case, frequent DB queries and call updates will degrade the system performance. To solve these problems, more efficient mobility management method is needed. We propose efficient mobile management using DLR method, which is to reduce location traffics for smaller cell and frequent terminal moving in boundary cell in IMT-2000 networks.

2 The Proposed Scheme

Each VLR acts as a LR and has a given fixed circle area around itself and IDs of VLRs which are included in its circle area. When a terminal powered on, The VLR which includes terminal becomes the LR of terminal and the terminal's latest location information is sent to the DLR when terminal changes its RA. This state is maintained as long as the terminal is located in the current k-circle area. When terminal moves to new VLR from the current LR's k-circle, the new VLR becomes the DLR of the terminal. By this manner, the k-circle of the terminal can be changed dynamically. This method can be performed easily by comparing VLR ids which current LR has with the VLR id where terminal moved. For example Fig. 1, suppose the first circle which consists of 7 VLRs and it has VLR id from VLR-1 to VLR-7 where current LR is VLR_1 and others are VLRs which are included in first circle area. The terminal is located in VLR_5 now. If the terminal moves to the new RA, VLR_6, CLR isn't changed. So, VLR_6 sends the terminal's new location information to VLR_1, current LR. If terminal moves to VLR_10, the current LR, VLRI, has no id of VLR_10. Thus the LR is changed. VLR_10 become the new DLR of terminal.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 649–654, 2005.

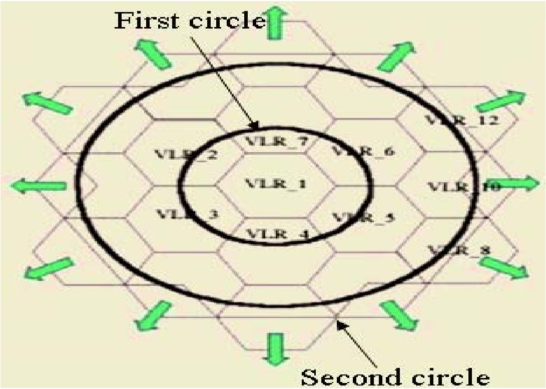


Fig. 1. Proposed DLR structure

In mobility management algorithm, this following shows the Pseudo-code for location registration and call tracking algorithm.

Algorithm Location Registration

```
{
Terminal's current LR id, VLR_xxx, received from old VLR;
Compare VLR_xxx with My_DLR_entry;
  If VLR_xxx exist in My_DLR_entry, then Send terminal_CURR_LOC to DLR;
    else { Write TID to MY_DLR_Area; //*belongs to a new DLR of the terminal
         Send terminal_CURR_LOC to HLR;
        Send REGCANC to VLR_xxx; } //*REGCANC is registration cancel message
    If call location update, then the terminal which moved to a new RA requests registration to the VLR of the new RA;.
The new VLR inquires the id of the terminal's current LR to the old VLR and the old VLR replies to the new VLR with ACK message including this information;
The new VLR calculate and determines whether the id of current LR exist in its VLR list of not;
end if hit, then after sending the location information of the terminal,
     the new VLR send a registration cancel message to the old VLR;
    else miss, then after transmitting location information of the terminal to HLR,
the new VLR transmits registration cancel message to old VLR and old CLR; }
```

Algorithm Call Tracking

```
DLR FIND( )
{
  Call to IMT-2000 user is detected at local switch;
   If called party is in the same RA, then return;
  else {
        Switch queries called party's HLR;
        Called party's HLR queries called party's current DLR, VLR_xxx;
        VLR_xxx returns called party's location to the calling switch; }}
```

3 Performance Analysis

3.1 IMT-2000 Mobility Model

To estimate the call cost, we assume a mobility model for IMT-2000 users. The direction of movement is uniformly distributed over [0, 2π]. The IMT-2000 users are uniformly populated with a density of ρ. The rate of RA crossing, R is (1/π)ρvL where the average velocity of users is v and RA boundary is of length L. In simulation, we assume the followings.

- RA size = $(7.575\text{km})^2 = 57.4\text{km}^2$
- Boundary length (L) = 30.3 km.
- Average call originate rate = Average call delivery rate = 1.4/*hr/terminal*
- Average terminal density (ρ) = 390/sq.km.
- User terminal/RA = $57.4 \times 390 = 22386$
- Average speed of user terminal (ν) = 5.6 km/hr.

Using above parameter indexes, we can calculate signaling traffic for registration, which occurs when user terminal moves to new RA.

$$R_{\text{reg.VLR}} = \frac{390 \times 30.3 \times 5.6}{3600\pi} = 5.85/\text{s} = R_{\text{Dereg.VLR}}$$

The following is the total registration message traffic which arrives to HLR per second.

$$R_{\text{reg.HLR}} = R_{\text{reg.VLR}\times} = \text{Total No. of RAs}$$

The number of queries which HLR must handle in call setup is following.

$$R_{\text{CallDeliv.HLR}} = \text{Call Rate per User} \times \text{No. of Users} = \frac{1.4 \times 22386}{3600} = 8.7/\text{s}$$

The switch certificates the terminal by querying the call which originated from the terminal to the serving VLR. This query generation rate is determined by the query rate generated in service switch point (SSP) area.

$$R_{\text{CallOrig.VLR}} = 8.7/\text{s}$$

The number of query which is needed for certification terminals is following.

$$R_{\text{CallDeliv.VLR}} = 8.7/\text{s}$$

Also, we adopt hexagon model as geometrical RA model. Let assume the DLR1 structure, which adopts first circle and exist two rings (0, 1): one VLR on the ring_ 0 and six VLRs on the ring_ 1. We can calculate the number of terminal which moves inside the first circle area and changes its DLRs. The number of terminal which changes its DLR is 3R, and the number of terminal which moves inside first circle area is 4R. To performance analysis, we define using signaling costs, SC i as follows.

SC1 = Message transmission cost from one VLR to another VLR through HLR $= 2\left(A_l + L + D + R + V_Q\right) + A_r + H_Q$

SC2 = Message transmission cost from one VLR to another VLR through RSTP $= 2\left(A_l + R + D + R + V_Q\right) + R$

SC3 = Message transmission cost from one VLR to another VLR through LSTP $= 2\left(A_l + V_Q\right) + L$

A_l = T_x / R_x message transmission cost between SSP and LISP on A-link.

D = T_x / R_x message transmission cost on D-link.

A_r = T_x / R_x message transmission cost between RSTP and SCP on A-link.

L = Message routing and its handing cost by LSTP.

R = Message routing and its handing cost by RSTP.

H_Q = HLR access cost , V_Q = VLR access cost.

We evaluate the performance according to the relative values of SC1, SC2, and SC3 which are needed for location registration and location tracking. Intuitively, we can assume SC3 $\leq$ SC2< SC1 or SC3 $\leq$ SC2 < < SC1.

Table 1. Cost parameter set for analysis

Set	SC3	SC2	SC1
1	1	1	1
2	1	2	3
3	1	2	4
4	1	3	5
5	1	3	6
6	1	4	7
7	1	4	8

3.2 Numerical Analysis

(1) Cost for Location Registration ($C_{\mathbf{Loc\ Reg}}$)

In the IS-95 scheme, whenever a terminal moves into anew RA, it is registered at the HLR. The registration cost is computed as follows.

$$C_{\mathrm{IS\text{-}95,Loc.Reg}} = 2\mathrm{SC1}$$

In the proposed scheme, registration cost is the sum of the costs where a terminal moves inside first circle area and changes its DLR. The cost for terminals which moving inside first circle area is a half of six VLRs which are located on ring_1 and one VLR on ring_0. All messages between VLRs are transferred through LSTP of RSTP. This takes place three cases.

$$C_{\mathrm{Pr}\,oposed\ \ Lo.\,\mathrm{Re}\,g.noCg} = \frac{16}{7} \times \mathrm{SC3}$$

$$C_{\mathrm{Pr}\,oposed\ \ Md.\,\mathrm{Re}\,g.noCg} = \frac{8}{7} \times \left(\mathrm{SC2} + \mathrm{SC3}\right)$$

$$C_{\mathrm{Pr}\,oposed\ \ \mathrm{Re}\,.\mathrm{Re}\,g.noCg} = \frac{16}{7} \times \mathrm{SC2}$$

The cost for terminals which move outside first circle area are half of six VLRs which are located on ring_1 and messages are transferred via LSTP or RSTP. This takes place two cases.

$$C_{\Pr oposed \quad Lo.\,\mathrm{Re}\,g.noCg} = \frac{6}{7} \times (SC1 \; + SC3)$$

$$C_{\Pr oposed \quad \mathrm{Re}\,.\,\mathrm{Re}\,g.noCg} = \frac{6}{7} \times (SC1 \; + \; SC2)$$

Total cost is the average of these cases which are worst case and best case.

$$C_{\Pr opoded \quad Total \;, worst} = (36\,SC\,1 + 78\,SC\;2 + 62\,SC\;3)\,/\,42$$

$$C_{\Pr oposed \quad Total \;, best} = (22SC3 \quad + 6SC1)/7$$

(2) Cost for Call Tracking ($C_{\text{Call Tr.}}$)

In the IS-95scheme, when a call originates from user terminal, HLR is queried first.

$$C_{IS-95.Call.Setup} = 2SC1$$

In the proposed scheme, to connect to the cached VLR, the signal must pass LSTP or RSTP. If the callee terminal is located in the remote position, the cost is sum of cost for message transmission via RSTP in case of cache hit in DLR1 and the cost for transmission via RSTP in case of cache miss and SCI1 in DLR1.

$$C_{\Pr oposed \;\; Hit, \mathrm{Re}\, mote} = (1 - 0.288) \times 2SC3 + \; 0.288 \times 2(SC1 + SC3)$$

If the callee terminal is located in the local position, the cost will be the sum of cache hit message transfer cost via LSTP and SC1 (tracking cost)

$$C_{\Pr oposed \;\; Hit, Local} = (1 - 0.288) \times 2SCISC2 + \; 0.288 \times 2(SC1 + SC2)$$

If there is no cached information of the callee terminal in the caller MSC, system tracks the terminal location by querying HLR and DLR1. The cost is similar to IS-95 scheme.

$$C_{\Pr oposed \;\; Cache.Miss} = 2SC1$$

In Fig. 2, we can see that proposed method has lower cost than IS-95 scheme, even though it is a worst case of proposed method provides mostly same cost as IS-95 scheme. The worst case takes places when ratios of six cases are same. In other word, it occurs when LSTP connected with very few RAs (i.e., less then three VLR/MSC). But we know that a LSTP's coverage is more than that of three RAs generally. The worst case of proposed scheme seldom occurs in actual networks. We know that the next generation wireless system will adopt smaller RA. It means that a LSTP will cover more wide registration area. We can see that the proposed scheme is more efficient than IS-95scheme.

In Fig. 3, proposed method has lower cost than IS-95 scheme, even though it is a worst case of proposed method provides mostly same cost as IS-95 scheme. The worst case takes places when ratios of six cases are same. In other word, it occurs when LSTP connected with very few RA. But we know that a LSTP's coverage is more than that of three RAs generally. The worst case of proposed scheme seldom occurs in actual networks. We know that the next generation wireless system will adopt smaller RA. It means that a LSTP will cover more wide RA.

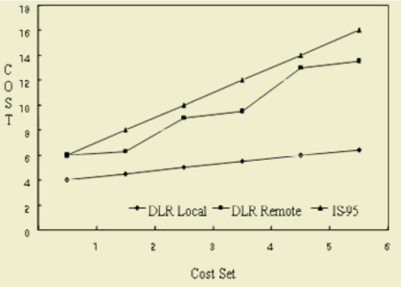


Fig. 2. Location registration cost

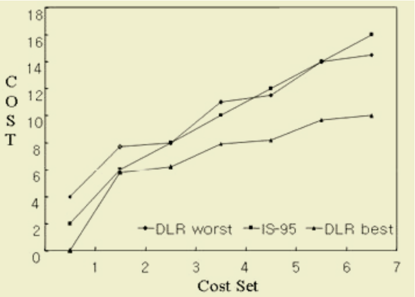


Fig. 3. Call tracking cost

4 Conclusion

We propose an auxiliary solution to improve the performance in previous schemes including IS-95 and GSM standard. The proposed DLR method is effective for smaller cell and more complex terminal moving pattern. This proposed scheme reduces location traffic according to change LR dynamically. As a result of cost evaluation, the more the VLRs in LSTP area are and the higher the call origination rate is, the performance is improved greater. The proposed scheme is much less affected by the terminals frequent RA crossings compared to the previous schemes including IS-95 and GSM standard.

References

1. A Bar-Noy and I. Kessler, " Tracking Mobile Users in Wireless Networks," IEEE INFOCOM'93, 1998.
2. R Jain and Y.B.Lin, "An auxiliary User Location Strategy Employing Forwarding Pointers to Reduce Network Impacts of PCS," ACM-Baltzer Journal of Wireless Network, Jul. 1999.
3. R.Jain, Y.B.Lin and S.Mohan, " A Caching Strategy to Reduce Network Impacts of PCS," IEEE Journal in Comm, vol 12, no. 8, Oct. 1998.
4. S.J.PARK, Dong Chun Lee and J.S Song, "Locality Based Location Tracking Using Virtually Hierarchical Link in Personal Communication Services," IEICE Trans. Com., VOL. Z81-B, NO. 9, Sept. 1999.
5. G, P, Pollin; cuel D, J, Goodmern, "Signaling System Performance Evaluation for Personal Communications," IEEE Trans. on Veh. Tech., May. 1998

SWSD: A P2P-Based System for Service Discovery from a Mobile Terminal

Darije Ramljak[1] and Maja Matijašević[2]

[1] IBM Croatia Ltd/Integrated Technology Services, Lastovska 2, HR-1000 Zagreb, Croatia
darije.ramljak@hr.ibm.com
[2] FER, University of Zagreb Unska 3, HR-10000 Zagreb, Croatia
maja.matijasevic@fer.hr

Abstract. This paper presents the architecture and implementation of the system named Semantic Web Services Discovery (SWSD), our proposed solution for service discovery from a mobile terminal. The key features of SWSD may be summarized as (1) semantic service description, based on DAML-S ontology for description of Web services, (2) service discovery, based on JXTA P2P technology, and, (3) mobile terminal support, based on JXTA for J2ME. The case study demonstrates a prototype implementation of the system for finding stock exchange information.

1 Introduction

With a variety of mobile Internet-enabled devices becoming more affordable and increasingly popular, there comes an opportunity for new revenues for content and service providers. Innovative, attractive services for mobile users range from messaging and "infotainment" to distributed games, multimedia streaming, and even augmented reality [8] on handheld PCs and mobile phones. Our work is motivated by the challenge to provide service discovery from a mobile terminal.

Even as new services are introduced, there is a need to devise a way in which a service can be described, discovered, and invoked over the network in a standard and interoperable way. Web services architecture [9] comprises a set of protocols and interfaces to support this concept, main ones being *Extensible Markup Language* (XML), *Simple Object Access Protocol* (SOAP), *Web Service Definition Language* (WSDL), and *Universal Description, Discovery and Integration* (UDDI). Unfortunately, UDDI is limited in terms of service description, as it does not address service semantics [2]. The need for addressing semantics in Web service discovery, and to facilitate "machine-processable description", has driven the development of other languages, notably, DAML-S [6] and its successor OWL-S [10], providing a semantic markup for Web Services. While these languages solve the problem of service description, a mobile user is faced with additional problem of service discovery when trying to access the service without prior knowledge of where the service resides. In such cases, the task of service discovery may be handed over to a peer-to-peer (P2P) network which provides a viable alternative to centralized registry services, such as UDDI.

In this paper, we describe a P2P based approach to discovering services from a mobile terminal, called Semantic Web Services Discovery (SWSD), based on JXTA

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 655–661, 2005.

for J2ME [1][7] P2P platform and DAML-S as a service description language. Our work is similar to the approach described in [4], but their approach is more geared towards combining multiple services to achieve complex tasks. Another interesting P2P service discovery approach, based on Gnutella P2P protocol, is described in [5].

The paper is organized as follows: Section 2 introduces the problem and motivation and Section 3 presents the architecture of the proposed SWSD system. Section 4 describes the implementation, and Section 5 presents the results, using as a case study a service to discover stock market information. Section 6 concludes the paper.

2 The Problem of Service Discovery by a Mobile User

In a service-rich and inherently distributed environment such as Internet, the mobile user should be able to access the service by requesting *what* the service should deliver (that is, give a semantic description), and do so without knowledge (or interest!) of *where* (which service provider's server) the service resides. From the providers' point of view, services (S*i*) may be offered by multiple providers (*A*, *B*), and/or they may be placed on, or replicated (*Si*, *Si'*), on multiple servers as shown in Fig. 1.

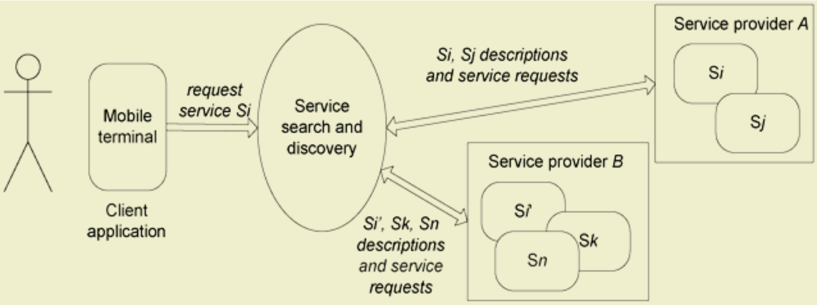

Mobile terminal
Client application
request service Si
Service search and discovery
Si, Sj descriptions and service requests
Service provider A
Si
Sj
Service provider B
Si'
Sk
Sn
Si', Sk, Sn descriptions and service requests

Fig. 1. The problem of service discovery

The process of service discovery should make it possible to match the semantic description of user's request against the semantic description of the registered service. Our approach is based on delegating the service request handling from a mobile terminal (with limited resources and capabilities) onto a P2P network.

3 The Proposed Semantic Web Services Discovery System

This section gives an overview of our proposed solution for service discovery from a mobile terminal, named Semantic Web Services Discovery (SWSD). The SWSD architecture is envisioned as comprising three layers:

- Application services,
- SWSD services,
- Core services.

We now describe each layer in more detail.

Application services layer consists of Internet services that the user is trying to find and use from the mobile terminal. The user view of the SWSD system is at this layer, while the two lower layers remain hidden. Services that the SWSD system provides in this layer are primarily based on the ontology in the particular service domain [3]. The applications that mobile terminals users utilize to search for information on the Internet may also reside in this layer. These applications can be very specific and tied to a particular domain, or, be general and provide information from a number of domains. Naturally, applications within a specific domain offer better search capabilities, but they lack ability to handle more diverse queries.

SWSD services layer includes services, mechanisms and protocols necessary for the operation of the SWSD system. In addition to services that are SWSD-specific, there are also core services of JXTA search platform (for distributed search), and mechanisms for semantic description of Web services (WSDL, UDDI). Main services that this layer offers to the Application services layer include *registration service*, *query service*, and *execution service*.

Registration service (Fig. 2) comprises mechanisms and protocols that required by SWSD system to register the services it provides.

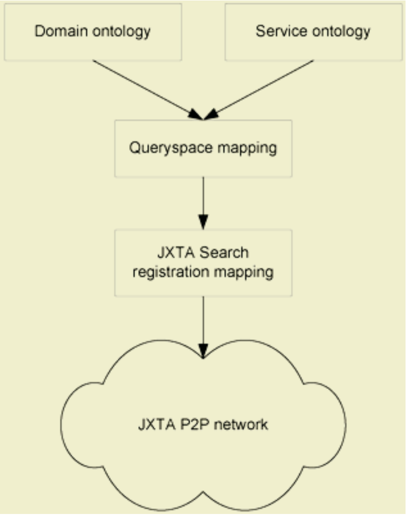


Fig. 2. Registration service

The service registration is executed at two logically separated levels where various mappings of data are performed. First level is semantic; where semantic description of the Web service is created using the mapping of DAML-S description and WSDL and UDDI, in order to store the semantic Web service description in the UDDI registries. After the service has been semantically described, the mapping to the P2P Search registration message is performed on the second, syntactic level where the service is registered in the JXTA Search P2P network. The mapping process transforms the XML received from the DAML-S mapping to the JXTA registration XML message, subsequently sent to the JXTA network. This message effectively registers the provider of the service in the *JXTA Search* P2P network.

Query service (Fig. 3) is the most important and by far the most complex service in the SWSD services layer. Clients use this service to send the query messages in the

SWSD system in order to find particular information. The clients must be registered with the system in order to use the SWSD service. The query service is also performed at two levels. At a higher level, query processing is performed semantically. The peer uses different matching mechanisms to compare information from the query with those found in its registries. The matching may result in responding to a query, or, forwarding the query to other peers in the network. These mechanisms use the semantic descriptions and compare the Web services metadata. At a lower level, processing of JXTA Search queries is performed. XML messages from the higher layer are translated into JXTA Search query messages and are routed using JXTA Search query routing mechanisms to the appropriate peers where the queries are processed. The equivalent translation process is performed once the response message is created.

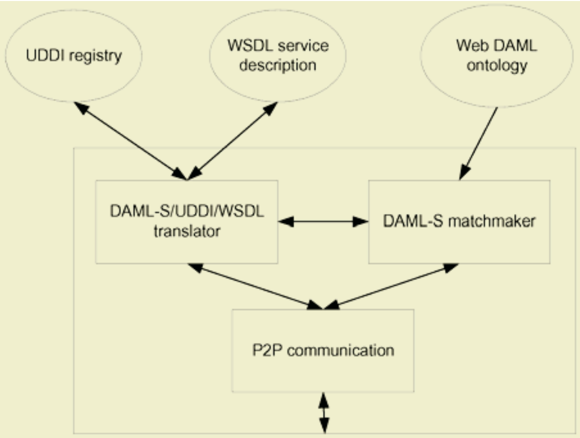


Fig. 3. Query service

Execution service may represent a standard invocation of a Web service, or, it may be routed through the SWSD system if it is performed right after query processing and the response message contains the information searched for.

Core services layer consists of services such as security, authentication and authorization. The security when communication between peers is implemented by using Transport Layer Security (TLS).

4 Implementation

We now describe the implementation of the SWSD system and its use to find particular stock price information from a mobile terminal as a case study. SWSD system implementation is comprised of the following modules, as shown in Fig. 4:

- Module for transformation between DAML and JXTA Search queries/ responses
- JXTA Search provider, client and hub
- Module for communication to the information service
- Module for registration of DAML/JXTA Search service
- Client GUI application

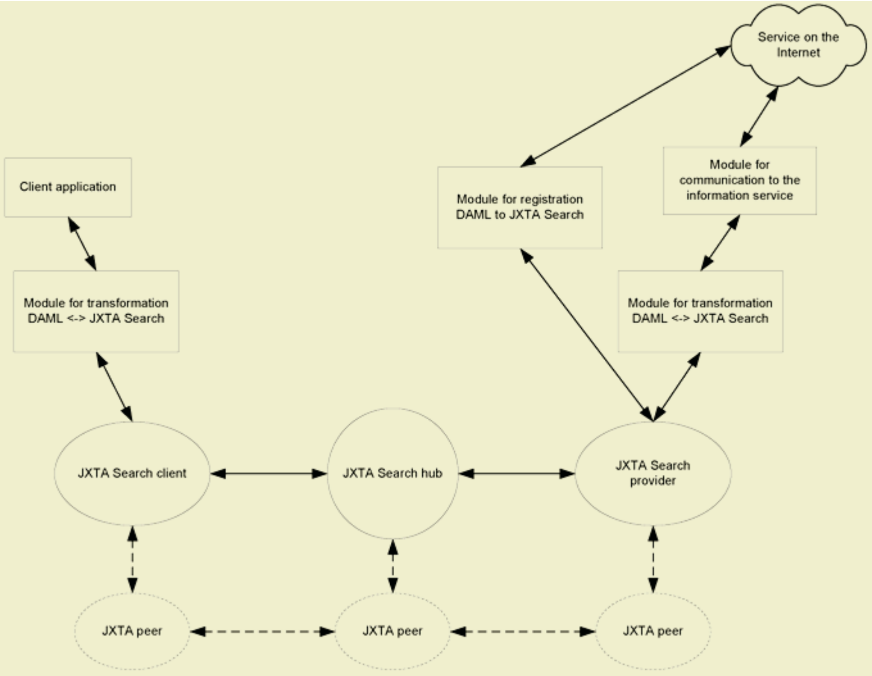

Service on the Internet
Client application
Module for registration DAML to JXTA Search
Module for communication to the information service
Module for transformation DAML <-> JXTA Search
Module for transformation DAML <-> JXTA Search
JXTA Search client
JXTA Search hub
JXTA Search provider
JXTA peer
JXTA peer
JXTA peer

Fig. 4. SWSD system implementation architecture

Module for transformation between DAML and JXTA Search queries/responses transforms the queries on the client side and the provider side. From the client side, JXTA search query can contain the arbitrary XML data, and this feature is used to integrate DAML query in the JXTA Search query message. The additional information is stored in some mandatory fields of JXTA Search message that help JXTA Search hubs to efficiently route messages to the peers that registered themselves as providers in SWSD system and a particular domain (in the case study, stock information).

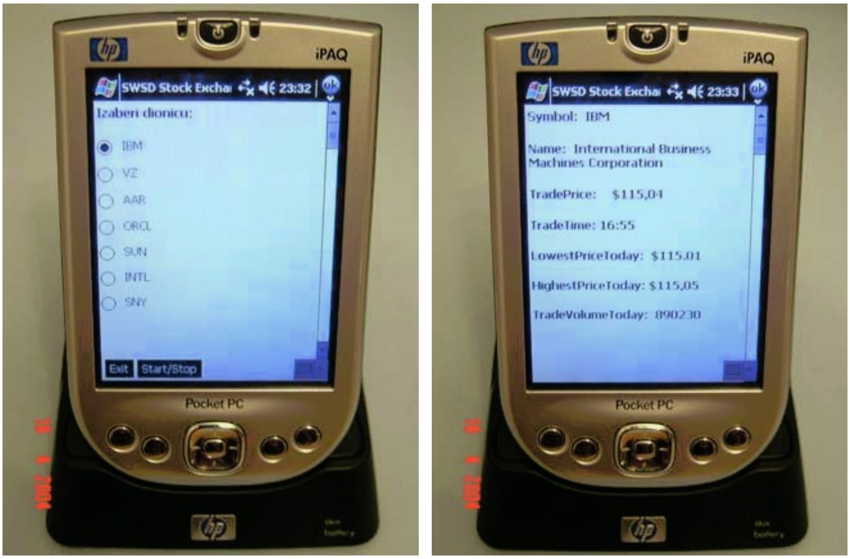
Module for the communication with the stock service is used to contact a Web service that offers stock information, and to prepare DAML response when the Web service sends back the information on the particular stock.

5 Results

To demonstrate the system operation, SWSD has been installed in the laboratory testbed. The JXTA Search client, provider, and hub (as standard elements of the JXTA Search P2P network) have been are installed on three personal computers (PCs). The client GUI application is a Java (J2ME) application that runs on iPAQ Personal Digital Assistant (PDA). The use case shows registering the stock information service with SWSD, and finding information by querying the system.

In order to make the service available to the P2P network, the stock registration service registers the service that publishes information about stock information to all potentially interested parties. Registration is performed by sending the JXTA Search registration message to the JXTA P2P network. This is accomplished using basic JXTA Search shell tools. After this, the stock information service is registered and the provider peer is ready to receive queries that JXTA Search hub will route to it.

The next step is finding stock information by sending a query to the SWSD system in order to retrieve information on a particular stock. The mobile terminal user uses the SWSD Stock Exchange client application to select the stock symbol for the stock he is interested in to start the discovery process. (Fig. 5, left).



Izaberi dionicu:

● IBM
○ VZ
○ AAR
○ ORCL
○ SUN
○ INTL
○ SNY

Symbol: IBM

Name: International Business Machines Corporation

TradePrice: $115,04

TradeTime: 16:55

LowestPriceToday: $115.01

HighestPriceToday: $115,05

TradeVolumeToday: 890230

Fig. 5. Client GUI application on mobile terminal

The selected stock exchange symbol is set as an attribute of DAML query and sent to the module of DAML/JXTA Search transformation. DAML query is transformed into the JXTA Search query and routed to the first available JXTA Search hub. Hub looks up its service registration tables and finds the peer that was registered as a provider of the stock information service and routes the query message. On the provider peer, the DAML query message is extracted from the JXTA Search message. Module for the communication with the information service extracts key elements from the DAML message and selects the Web service that was registered as a provider of stock information. After the response is received, it is routed all the way back to the client. The GUI on the mobile terminal shows the result of the search process (Fig 5, right).

6 Conclusion

This paper presents the application of P2P technology and semantic services to solve the problem of Internet services discovery from a mobile terminal in a system called SWSD. The main advantages of the SWSD system architecture are the ability to discover services by using semantic service descriptions, and the P2P network for distributing the load during the search process. A layered architecture allows seamless upgrades of system functionality. The use of Java-based JXTA platform makes it independent of potential changes in communication protocols between peers. The most significant drawback of the SWSD system is that all the base technologies used in it, JXTA for J2ME and DAML-S, are rather recent and subject to ongoing development. (For example, as mentioned earlier, in the course of this work, DAML-S was succeeded by OWL-S.). To keep up with standardization efforts and developments, additional work might be required. Other possible improvements include improving security and implementing role based authorization.

References

1. Arora, A., Haywood, C., Pabla, K.S.: JXTA for J2ME^{TM}- Extending the Reach of Wireless With JXTA Technology, Sun Microsystems, Inc. White paper (2002) 5pp.
2. McIlraith, S. and Martin D.: Bringing Semantics to Web Services, IEEE Intelligent Systems, Vol. 18, No. 1 (2003) 90–93
3. Paolucci, M., Kawamura, T., Payne, T.R., Sycara, K.: Semantic Matching of Web Services Capabilities. Proc. of the 1st Int. Semantic Web Conference (2002)
4. Sheshagiri, M., Sadeh, N., Gandon, F.: Using Semantic Web Services for Context-Aware Mobile Applications, Proc. of MobiSys2004 Workshop on Context Awareness, Boston, MA, USA (2004)
5. Paolucci, M., Sycara, K., Nishimura, T., Srinivasan, N.: Using DAML-S for P2P Discovery, Proc. of the 1st Int. Conf. on Web Services ICWS'03, Las Vegas, NV, USA (2003) 203-207 .
6. Ankolenkar, A., et al.: DAML-S: Web Service Description for the Semantic Web, Proc. of the First International Semantic Web Conference (ISWC), LNCS Springer-Verlag, Berlin, Germany (2002)
7. Gong, L.: JXTA: A Network Programming Environment, IEEE Internet Computing, Vol. 5, No. 3 (2001) 88–95
8. Christopoulos, C.: Mobile augmented reality (MAR) and virtual reality. Wireless World Research Forum (WWRF): The book of visions 2001. (2001) [http://www.wireless-world-research.org/]
9. Booth, D., et al. (Eds): Web Services Architecture, W3C Working Group Note, Feb. 2004, (2004) [Available: http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/]
10. Martin, D., et al. (Ed.): OWL-S: Semantic Markup for Web Services, Nov. 2004, (2004) [Available: http://www.daml.org/services/owl-s/1.1/]

An Efficient Eye Location Using Context-Aware Binarization Method

Jo Nam Jung, Mi Young Nam, and Phill Kyu Rhee

Dept. Of Computer Science & Information Engineering. Inha University,
253 Yong-Hyun dong, Incheon, South Korea
jjn10@korea.com, rera@im.inha.ac.kr, pkrhee@inha.ac.kr

Abstract. Face recognition system needs the method of locating facial components like eyes, mouth, etc, for extracting features used in recognition process. In this paper we propose a method of locating eyes using context-aware binarization. The proposed method consists of binarization, connected region segmentation, eye candidate area extraction by heuristic rules that use geometric information, eye candidate pair detection, and eye area pair determining by ranking method. Binarization plays an important role in this system that converts a source image to a binary image suitable for locating eyes. We consider edge detection based and segmentation based binarization methods. However, each method cannot be a solution in general environment because these are influenced by the factors such as light direction, contrast, brightness, and spectral composition. We propose a hybrid binarization using the concept of illumination context–awareness that mixes two binarization methods in general environment. We apply this methodology to eye location, and we achieved encouraging experiment results in general environment.

1 Introduction

The communicative power of the face makes it a focus of attention during social interaction. To fully understand the subtlety and informativeness of the face, considering the complexity of the movements involved, one must study face perception and the related information processing. For this reason face perception and face processing have become major topics of research by cognitive scientists, sociologists and most recently by researchers in computer vision and computer graphics [6],[7],[10]. The automatic processing of human face by computer will be significant step towards developing an effective HCI(Human computer interface)[1],[8],[9]. Eye location with the information of face region consists of largely preprocessing, binarization, connected region segmentation, detecting of candidates for eye region and eye pair using heuristic rules based on geometric information and determining eye pair by Ranker. Binarization method that converts an original image to a binary image suitable for eye location is considered as edge detection and segmentation. However, each method is dependent on the environmental factors such as light direction, contrast, brightness, and spectral composition. For solving this problem we propose a new method of mixing two binarization method. We introduce a concept of the context-aware binarization for solving this problem. The changes of illumination environment can be detected by analyzing the input images. We assume that the illumination environment changes continuously. We apply this methodology to eye location.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 662–669, 2005.

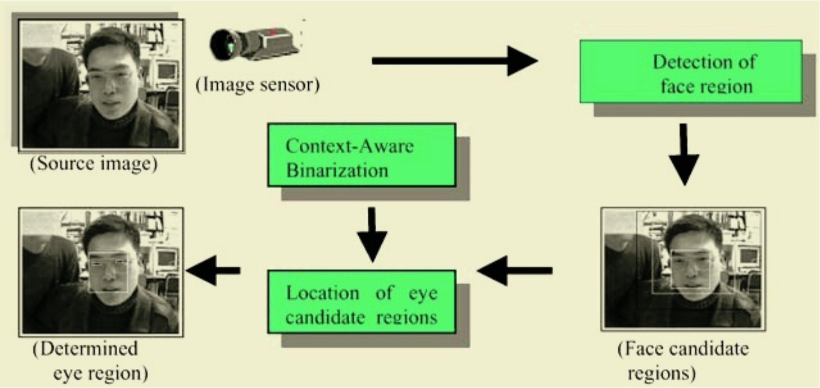


Fig. 1. Block diagram of eye locating process

2 Binarization

Chapter 2 roughly explains edge detection based and segmentation based binarization.

2.1 Edge Detection Based Binarization

Image edges are defined as local variations of image intensity. This variation is gotten by edge detector operators. The image gradient $\nabla f(x,y)$ and magnitude of image gradient $e(x,y)$ is given by[4],

$$\nabla f(x,y) = \left[\frac{\partial f}{\partial x} \frac{\partial f}{\partial y} \right]^T = \Delta [f_x, f_y]^T \qquad (1)$$
$$e(x,y) = \sqrt{f^2{}_x(x,y) + f^2{}_y(x,y)_x}$$

Edge Image is obtained by an edge detector using *e(x,y)*. This image carries information about the edge magnitude. If the edge detector output is large, a local edge is present. Otherwise this pixel corresponds to background. We get a better image suitable for locating facial components using the Sobel edge detector [11]. This provides good performance and is relatively insensitive to noise. However, the basic edge detection method is very dependent on threshold *T* that largely influenced by environmental factors, too. Therefore, if we use threshold as a fixed value we cannot obtain a good image because we always experiment in different environments. For solving this problem we experiment a method getting a threshold *T* using the statistical information of the edge magnitude and the property that the larger the edge magnitude is, the higher possibility of being edge is.

2.2 Image Segmentation Based Binarization

The shape of an object can be described in terms of the eye region it occupies. This region can be obtained by image segmentation. Image regions are expected to have homogeneous characteristics (e.g. intensity, texture) that are different in each region. These characteristics form the feature vectors used to discriminate one region from the other. The features are employed during the segmentation procedure in the rules that check region homogeneity [4]. This following equation is the definition of segmentation adopted in this paper.

$$E(x,y) = \begin{matrix} 1 & \textit{if} \quad e(x,y) \geq T \\ 0 & \textit{otherwise} \end{matrix} \tag{2}$$

where T is threshold whose choice can be based on the image histogram.

Image segmentation produces a better performance than edge detection in case that an original image has low quality or contains a face wearing spectacles. Image segmentation as our binarization method uses the property that the major facial components are darker than the gray color of the skin in case of the yellow race. Using this property and the statistical information of image intensity we can get the segmentation threshold *T* enough to obtain the pixels consisted of eye regions. Next, segmentation produces the binary image by making use of *T* as segmentation threshold. This segmentation method has the following property. The larger a common region between the region of face candidate and the region of real face is, the better performance is. The reason is that the smaller the common region is, the larger the region of background or the hair of the head is and thus the higher a probability of including the region darker than eye regions is. Therefore, this method is much dependent on the accuracy of the region of face candidate.

3 Locating of Eye Regions

Fig.2 roughly describes the model that we propose for locating eye regions. Eye location begins with inputting each face candidate regions into eye location process given in Fig.2. However, if image is captured in weak illumination, generally face candidate region becomes a small intensity region, i.e. Face candidate region's image contrast become poor and the subjective image quality become low. For enhancing image quality, histogram equalization that modifies its histogram performs in face candidate regions[5].

We found that binarization processed by edge detection algorithm is efficient when the candidate region is a dark image and binarization by segmentation algorithm is efficient when the candidate region is normal or bright from experiment. Input face candidate region is analyzed using neural network in the context awareness module. The binarization is performed differently according to the analyzed illumination condition. If the image illumination condition is dark, an input face candidate region is binarized by the edge detection method. If the condition is normal or bright, it is done by the segmentation method.

3.1 Face Candidate Regions

Eye locating is executed in face candidate regions detecting by mosaic and back propagation neural network[3]. However, face candidate region exists more than one. This means that face candidate region is the completely wrong region and whether it can contain a pair of eyes or not.

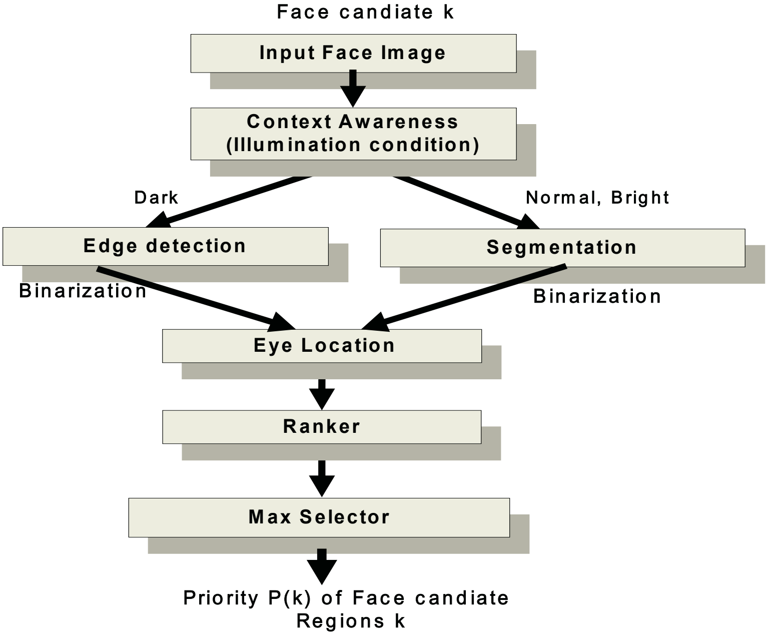


Fig. 2. The block diagram of eye location

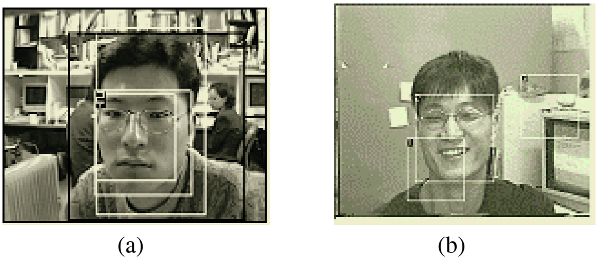

(a) (b)

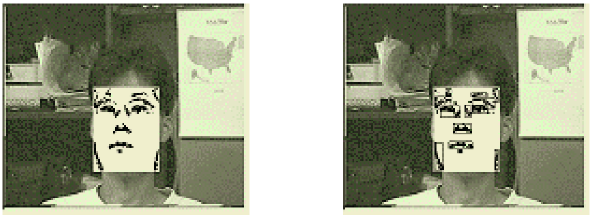
Fig. 3. Examples of face candidate regions

3.2 Binarization and Connected Region Segmentation

Using the binarization method explained in section 2 and region connectedness that examine the connected state of between a pixel and its 8-neigbourhood pixel, we obtain many regions that are connected and isolated each other[2].

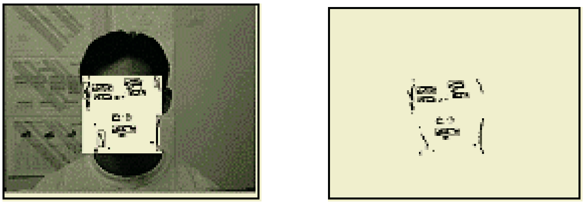
3.3 Detection of Candidates of Eye and a Pair of Eyes

Candidates of eye and a pair of eyes are detected by heuristic rules based on eye's geometrical information in face. Eye candidate’s width has much correlation with face candidate's height. However, because the height of eye candidate has an extreme changeability on condition of blinking one's eyes, it is difficult to set a correlation between eye candidate’s height and face candidate region. Eye detection has a tendency to acquire the boundary of eye. However, image segmentation have a more concern in pupil regions than the boundary of eyes, so the connected regions by labeling have a tendency to become more smaller and more nearer a square than ones obtained by edge detection. Therefore, rules used in eye location differ as to binarization methods. The example of detecting candidates of eye region is shown in Fig. 5.



(a) Binarization image (b) Connected region segmentation

Fig. 4. Example of connected region segmentation



(a) Connected region segmentation (b) Candidates of eye region

Fig. 5. Example of detecting candidates of eye region

Once eye candidate regions are detected, candidates of a pair of eyes are detected. First, because two eyes locate in a similar position by y -axis, eye candidate regions are sorted by y -axis. Candidates of a pair of eyes satisfy the following rules- the gradient of two eyes, comparison between two eye's size and distance between two eyes. These rules also differ as to binarization method similar to the rule of eye candidate regions.

3.4 Ranker

After eye location process, Ranker calculates an energy of each candidate of a pair of eyes detecting in each face region F_{k}. Each energy obtained by Ranker inputs to *Max* and it selects a max value among those. A candidate of a pair of eyes whose energy is equal to this selected value become a pair of eyes.

Fig.6 shows a template of a pair of eyes that are represented with terms of the size, shape, positional information of eyes, eyebrows and mouth. Using this template, Ranker calculates an energy of a pair of eyes.

Eq(3) is an equation representing a template given in Fig.6. E_L, E_R, E_M which are the energy, using terms of image intensity, edge magnitude, the region information obtained

$$\begin{aligned} E = {} & k_1 E_L + k_2 E_R + k_3 E_M + k_4 (E_{LR} + E_{LM} + E_{RM} \\ & + k_5 (E_{LL} + E_{RR}) \end{aligned} \tag{3}$$

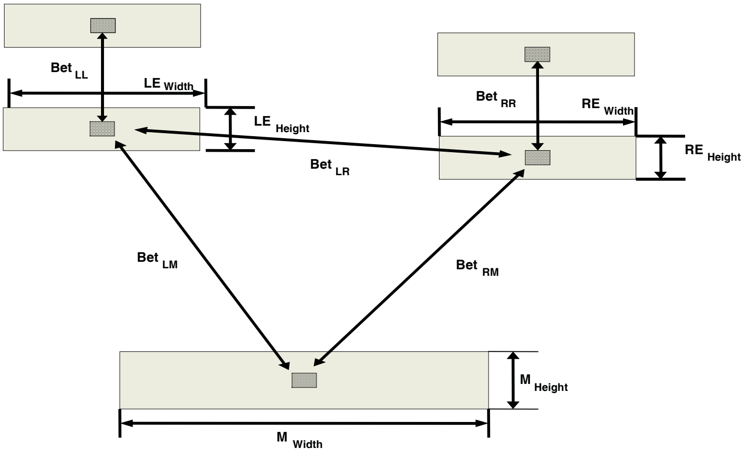


Fig. 6. The template of a pair of eyes

by connected region segmentation and so forth, calculates in the left eye, right eye, mouth respectively. In addition, E_{LR}, E_{LM} and E_{RM} are the energy between two eyes, the left eye and the mouth, the right eye and the mouth, respectively. Finally, E_{LL} is the energy between the left eye and the left eyebrow, and E_{RR} is the energy between the right eye and the right eyebrow.

3.5 Determination a Pair of Eyes

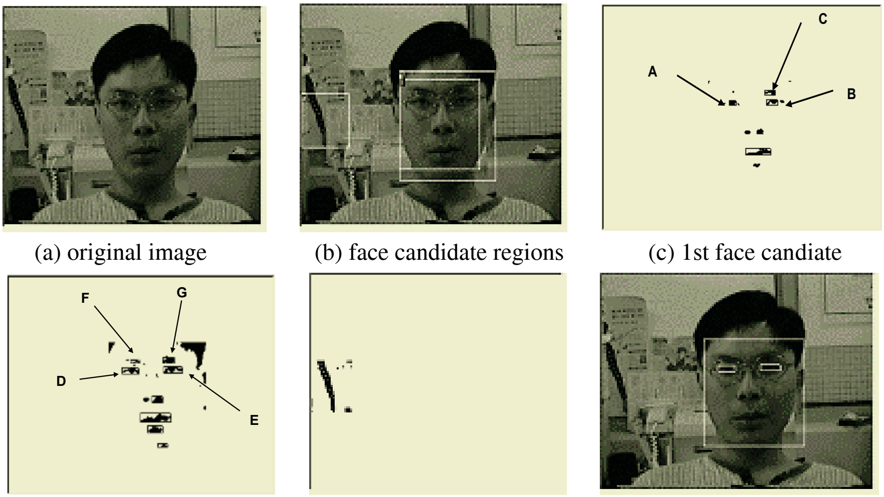
There can exist more than one candidate for a pair of eyes obtained in each candidate face region. Now, determined eyes have a maximum value of energy by Ranker. Therefore if there exist *m* candidates for a pair of eyes obtained in a face candidate region *Fi*, each candidate for a pair of eyes is expressed as Eye^{i}_{k}, $1<=i<=m$ and its energy calculated by Ranker is define as $E(Eye^{i}_{k})$. Therefore, the energy of a determined pair of eyes in face candidate region *Fi* are defined as below,

$$E(k) \;=\; \underset{1\le i\le m}{MAX}\left\{P(Eye^{i}_{k})\right\} \tag{4}$$

and finally, the determined eyes of this facial image should satisfy Eq(5).

$$\underset{1\le k\le m}{MAX}\left\{E(k)\right\} \tag{5}$$

Fig.7 shows the examples of determination of eye regions. Three face candidates are displayed in Fig.7(b). The left top corner of each face region has face candidate's rank. Fig.7 (c), (d), (e) are the results of executing the binarization, connected region segmentation and conditioning the rule that will be satisfied by a candidate for a pair of eyes. From the 1st face candidate given in Fig.7(c) two candidate for a pair of eyes *(A,B)*, *(A,C)* was detected and from the 2nd face candidate *(D,E), (D,G), (F,E), (F,G)*. However, there's no candidate for a pair of eyes from 3rd face candidate region. With the values of the energy calculated from the above seven candidates, we determine 2nd face candidate's *(D,E)* as eye regions because of it having maximum value among them.



(a) original image (b) face candidate regions (c) 1st face candiate

(d) 2nd face candidate (e) 3rd face candidate (f) determined eye regions

Fig. 7. Examples of determination of eye regions

4 Experiments and Conclusions

Our proposed method of locating eye regions was developed with MFC(Microsoft foundation class) in Pentium Ⅳ 2.4GHz CPU PC having Windows XP as operating system. Compiler used was Visual C++ 6.0. Experimental images were captured from image sensor with 320×240 size and 256 gray levels and were collected totally 620 frames from 41 persons(see Table 1). Table 1 shows the successful rate of location of eye regions. The successful rate is given by

$$\text{Successful rate} = {}^{(\text{Number of success})}\!\!\diagup_{(\text{Number of images})} \tag{6}$$

Table 1. Comparing results of different methods of eye location

Glasses	Method of Binarization	Number of Images	Number of success	Successful rate
Not wear glasses	E	398	391	98.24%
	S		372	93.47%
	E+S		391	98.24%
Wear glasses	E	222	167	75.23%
	S		200	90.09%
	E+S		208	93.69%
Sum	E	620	558	90.00%
	S		572	92.26%
	E+S		599	96.61%

This successful rate of eye locations largely depends on the environmental factors image quality, illumination, glasses and hair of head and so forth. In this experiments, only the factor of glasses was considered and experimental images were classified by the factor of glasses. These classified images were separately experimented with different methods edge(E), segmentation(S) and combined method(E+S) and were compared with each result.

According to the above result using combined method was superior to single method. In a stable illumination, in case not wearing glasses we could get successful rate using only edge method. However, if glasses were worn, successful rate of edge method was lower than that of segmentation method and at the same time especially in case black glasses worn frame edge method made the rate of success remarkably lowered. Consequently, using a single method is suitable for a particular environment. Nevertheless, because the environment getting images isn't always stable in the real world, if we can't select appropriate method we can't get a good result. So, in this paper we combined two binarization method for locating eye regions and can get a good result in a general environment as well as a particular one.

In the future's experiments, we will have experimented on locating eye regions in the more real world's environments an object near eyes and the reflection of a beam of light off a glasses and so forth.

References

1. P.Ekman, T.Huang, T.Sejnowski, and J. Hager (Editors): Final Report to NSF of the planning Workshop on Facial Expression Understanding. Technical report, National Sciencs Foundation, Human Interaction Lab., UCSF, CA 94143 (1993)
2. R. L. Lumina, G. Shapiro, and O. Zuniga: A New Connected Components Algorithm for Virtual Memory Computers. Computer Vision, Graphics, and Image Processing, Vol. 22. (1983) 287-300
3. Jung-Il Choi, Su-hwan Kim, and Phill-Kyu Rhee: Face Detection based on Multi-level Neural Network. Proceedings of the 8th KIPS Fall Conference, Korea Information Processing Society. (1997) 565-570
4. I. Pitas: Digital image processing algorithms. Prentice Hall (1992)
5. R. Klette, P.Zamperoni: Handbook of image processing operators. John Wiley&Sons (1996)
6. Paperno D, Semyonov D: “A new method for eye location tracking”. Record 7 of 11 from IEEE Transactions On Biomedical Engineering, no.10 (2003) 1174-1179.
7. Zhou ZH, Geng X: Projection functions for eye detection, Pattern Recognition 37(5), no.(5) (2004) 1049-1056
8. Bin Chen, Zhi-Qiang Liu, and Xiang-Hua Zhu: Eye Location In Human Face Images Using Fuzzy Integral. IEEE Trans. On Proceedings of the Second International Conference on Machine Learning and Cybernetics. Vol.14. no.2-5(11) (2003) 2500-2505
9. Li Weijun, Xu jian, and Wang Shoujue: A Fast Eye Location Algorithm Based on Geometric Complexity. Proceeding of the 5th World Congress in Intelligent Control and Automation, Hangzhou, P.R. China, Vol. 5. no. 6 (2004) 4105-4107
10. Junu Park, Jo Nam Jung, and Phill Kyu Rhee: An Effective Eye Location for face Recognition. Image Processing and Image Understanding Workshop on Ramada Plaza Jeju Hotel. The Institute of Electronics Engineers of Korea, Korea Information Science Society, Korean Institute of Communication Sciences, Vol. 17. no.21(1) (2005) 253-258
11. Qu Ying-Dong, Cui Cheng-Song, Chen San-Ben and Li Jin-Quan: “A fast subpixel edge detection method using Sobel-Zernike moments operator”, Image and Vision Computing, China, Vol. 23 Issue 1, no.1(1) (2005) 11-17

Design and Implementation of Context-Awareness Processor for Multiple Instructions in Mobile Internet Environment

Seungwon Na[1] and Gu-Min Jeong[2]

[1] SK Telecom CO., LTD. Terminal R&D Center
11, Euljiro 2-Ga, Chung-Gu, Seoul 100-999, Korea
nasw@dgu.ac.kr

[2] School of Electrical Engineering, Kookmin University
86-1 Jungnung-Dong Sungbuk-Gu, Seoul, 136-702, Korea
gm1004@kookmin.ac.kr

Abstract. The usage of cellular phone has been expanded as wireless internet device recently. But its popular use is limited by some of reasons. The main reason is unshared data between data base servers. So, users can only access data that they know the link or location. In this paper, MRSS (Mobile Resource Sharing System) is presented to overcome the local limits of data and enhance data sharing methods. MRSS offers better service by using Context-Awareness Processor that transforms one instruction to many related instructions. MRSS is going to activate wireless internet service and provides a ubiquitous computing environment to customers using the mobile Internet.

1 Introduction

Cellular phone provides portability to users and is spreading throughout our lives. Cellular phone was used for wireless phone calls at early times, but its usage has been expended as a wireless Internet device recently [1]. But practical use of the mobile Internet is limited by unshared resource. There are 30 million Korean mobile Internet users in 2004. But only 5% of them accessed wireless internet for more than 24 hours per month [2]. It means there was only quantitative improvement, but it is not qualitative one. The reason of this comes from the structure of mobile Internet network that provides only limited information [3].

In this paper, MRSS is presented to concretize mobile resource sharing by overcoming location restriction and expanding mutual connectivity of resources. The core of MRSS is Context-Awareness Processor and it creates multiple instructions from a single instruction. The implementation of MRSS is based on Context-Awareness of Ubiquitous Computing Environment. This paper is considering mobile Internet that works in cellular phones, but not laptop computers that have Internet access devices such as wireless LAN and Bluetooth.

2 Related Works

2.1 Ubiquitous Computing

Mark Weiser said "Ubiquitous Computing has as its goal the enhancing computer use by making many computers available throughout the physical environment, but mak-

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 670–676, 2005.

ing them effectively invisible to the user" [4]. Ubiquitous Computing is consisted of three components such as Pervasive Sensing, Context-Awareness and Context-Management [5,6]. First, Pervasive Sensing means notification of information occurrence from users and external environment. Second, Context-Awareness means awareness of requested phrases and contexts by processing and sharing information acquired from scattered sensors and other networks. Third, Context-Management means management of external network and database server.

2.2 Combination of Context-Awareness and Mobile Computing

There are two ways to construct Ubiquitous Computing environment. One of them is spreading USN (Ubiquitous Sense Network) chips at all of the places. The other is using portable devices. While the first way means quantitative expansion with physical expansion of resources and it is concept of Pervasive Computing that Mark Weiser mentioned. The second way means expansion of connectivity with logical expansion of resources. Improvement of Mobile Computing promotes the second way and shows Ubiquitous Computing up again. Mobile Computing has to be combined to a system built effectively and responds fast, so it gives connectivity of location in Ubiquitous Computing Environment [7, 8].

2.3 Context-Awareness and Processing Technology of Natural Language

Natural Language Processing is a computer technology that processes human language. Generally, it consists of 4 steps such as Morphological Analysis, Syntactic Analysis, Semantic Analysis and Pragmatic Analysis [9].

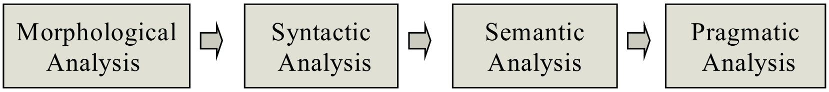


Fig. 1. 4 Steps of Processing Natural Language

First, Morphological Analysis is a step of separating sentences inputted as a token. Second, Syntactic Analysis is a step of deriving the normal structure of sentences by using information that is from Morphological Analysis. Third, Semantic Analysis is a step of interpreting the meaning of sentences by the result derived from Syntactic Analysis. This step determines whether it is semantically proper or not. Fourth, Pragmatic Analysis is a step of changing sentences to other formats that users want by mutual relationship of real world. In this paper, Natural language Processing is applied to Context-Awareness Processor.

3 Design of Mobile Resource Sharing System

3.1 Introduction of System

MRSS is consisted of RSC (Resource Sharing Client) in terminal area and RSS (Resource Sharing Server) in the server area. Fig. 2 shows the overall structure of the system.

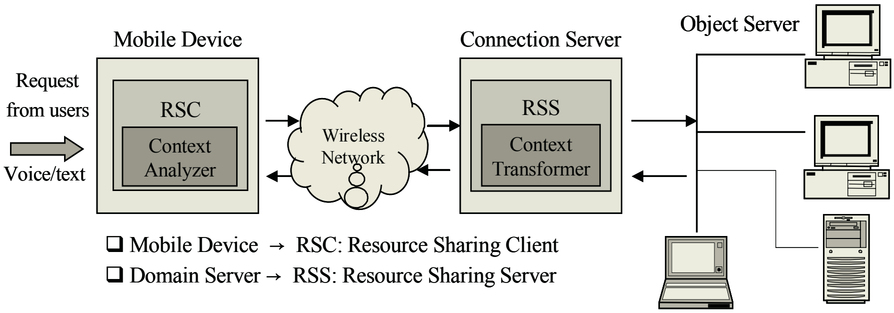


Fig. 2. Structure of Mobile Resource Sharing System

The core engine of MRSS can be built by constructing the Context-Awareness Processor. The implementation of MRSS is originated by Ubiquitous Computing and the system is physically divided by two parts. Fig.3. shows the overall structure of Context-Awareness Processor.

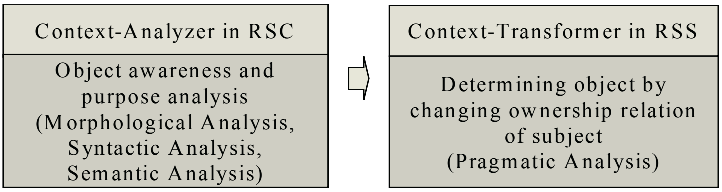

Context-Analyzer in RSC
Object awareness and purpose analysis (Morphological Analysis, Syntactic Analysis, Semantic Analysis)
Context-Transformer in RSS
Determining object by changing ownership relation of subject (Pragmatic Analysis)

Fig. 3. Functions of Context-Awareness Processor

3.2 Design of RSC

RSC creates instructions by user request and it automatically analyzes information in mobile devices. RSC has to be run on the operating system of devices and supports compatible data structures by middleware. Fig.4 shows this as well. RSC consists of 3 modules and shows the structure of RSC.

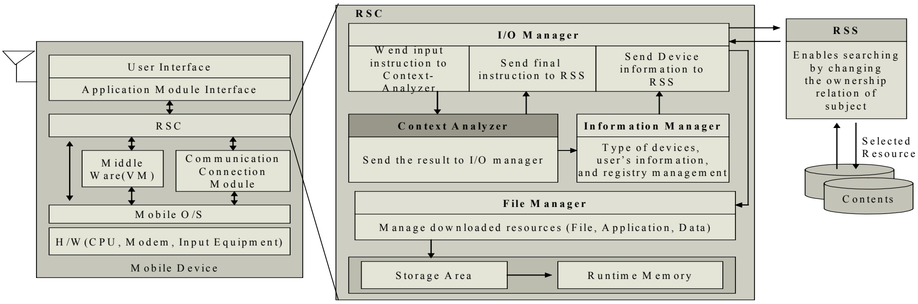

RSC
I/O Manager
Wend input instruction to Context-Analyzer
Send final instruction to RSS
Send Device information to RSS
Context Analyzer
Send the result to I/O manager
Information Manager
Type of devices, user's information, and registry management
File Manager
Manage downloaded resources (File, Application, Data)
Storage Area
Runtime Memory
RSS
Enables searching by changing the ownership relation of subject
Selected Resource
Contents
User Interface
Application Module Interface
RSC
Middle Ware(VM)
Communication Connection Module
Mobile O/S
H/W(CPU, Modem, Input Equipment)
Mobile Device

Fig. 4. Client Module and Process Resource Sharing

Context Analyzer in RSC accepts user enquiries and generates instruction by awardable word, case and purpose. And it requests to RSS for data. Other modules are only supporting Context Analyzer. Fig.5. shows the structure of Context Analyzer.

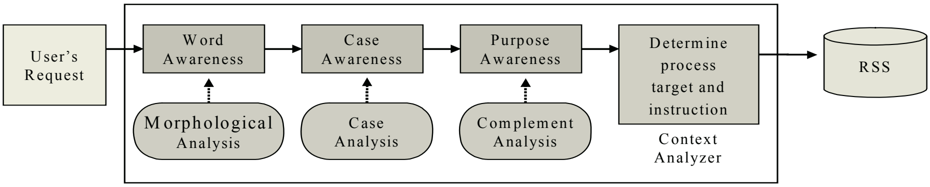


Fig. 5. Structure of Context Analyzer

For instance, "Print graduation photo of my wife" can be analyzed by Context-Analyzer as follows. First, every single word in the sentence is categorized as each morpheme. Second, they are categorized as cases. Third, the purpose of whole sentence can be derived by interpret verve. Finally, RSC can send determined information to RSS.

3.3 Design of RSS

RSS multiplies and executes the instruction by changing relation of the ownership. RSS consists of 3 parts – Context Transformer, Search Engine and File Manager. Fig.6. shows the structure.

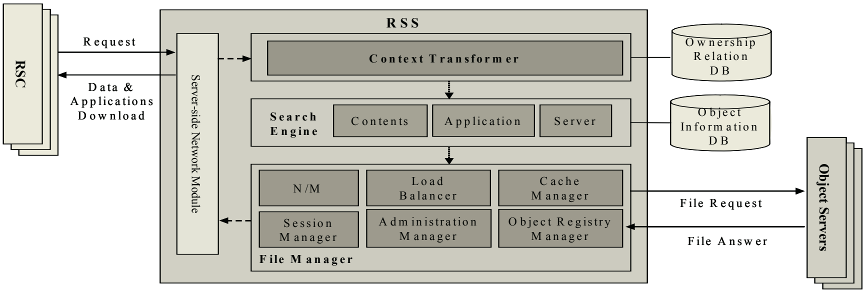


Fig. 6. Structure of RSS

First, Context Transformer changes the instruction from RSC into the form that users want. Various forms of instruction are generated by changing ownership relation of subject. Second, Searching Engine searches objects that fit on final instruction determined by Context Transformer. Third, File manager manages input and output information, session and registry. It also manages network modules and load-balances for sequential process of multiple access of RSC. <Table 1> shows DB table that explains functions and features of RSS modules.

Based on design as above, a single instruction can be transformed into multiple instructions by the process of Context-Awareness and Context Transformer. The process of changing ownership by multiple requests is shown in Fig.7.

Table 1. Functions of RSS Modules and DB Table

RSS modules		Functions
Context Transformer		Generate various instructions changing ownership relation
Search Engine	Contents Search Engine	Search the real location of information from DB
	Application Search Engine	Search the application for the contents
	Server Search Engine	Search the liking information to server that has actual contents or application
File Manager	I/O Manager	Manage input and output of contents
	Session Manager	Manage session
	Load Balancer	Distribute the server process
	Cache Manager	Cache files for service to RSC
	Administration Manager	Supply functions for server managers
	Object Registry Manager	Register target object information of object server in RSS
	Network Module	Manage network for responding and requesting instruction from RSC

RSS Database	Functions
Ownership Relation DB	DB information of ownership relation
Object Information BD	Information that needs to searching servers

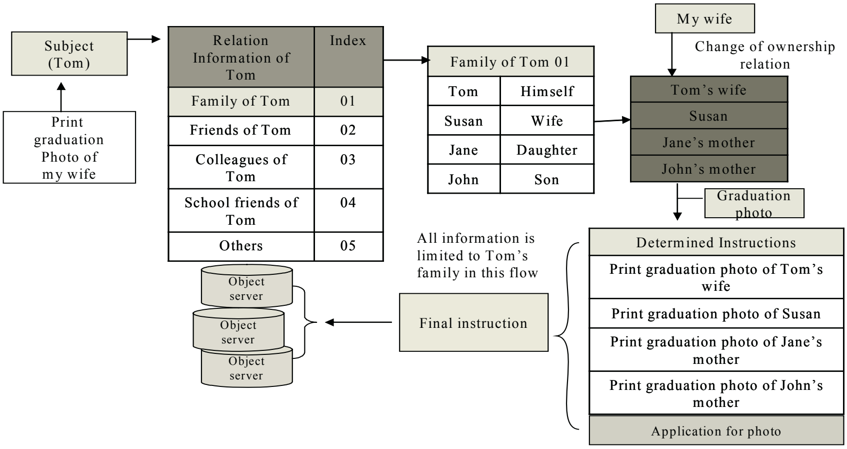


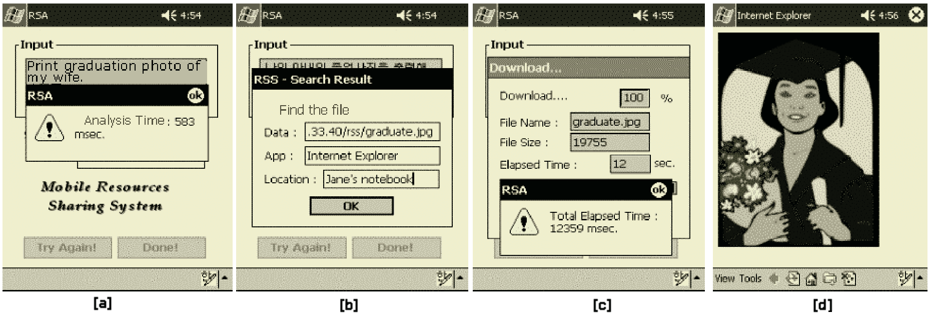
Fig. 7. Process of Changing Ownership Relation

As a conclusion of this chapter, RSS enables mobile resource sharing by magnifying connectivity of data scattered in many places.

4 Implementation and Result of Experiment

✓ Mobile Device: PDA (PPC 2002)
✓ Network Speed: CDMA 1x (144 Kbps/sec)
✓ Data format: photo (jpeg format), execution file
✓ Test Sentence: Print graduation photo of my wife

There are 4 steps for the experiment. They are shown Fig.8.



RSA 4:54

Input

Print graduation photo of my wife.

RSA ok

Analysis Time : 583 msec.

Mobile Resources Sharing System

Try Again! Done!

RSA 4:54

Input

RSS - Search Result

Find the file

Data : .33.40/rss/graduate.jpg

App : Internet Explorer

Location : Jane's notebook

OK

Try Again! Done!

RSA 4:55

Input

Download...

Download.... 100 %

File Name : graduate.jpg

File Size : 19755

Elapsed Time : 12 sec.

RSA ok

Total Elapsed Time : 12359 msec.

Internet Explorer 4:56

View Tools

Fig. 8. Process Steps of Context Analysis

[a] shows a screen shot that the system took 0.583 seconds to analyze Morpheme for testifying grammatical suitability in the Client. [b] shows the respond from servers that act on the instruction. It is the search result of object called from Jane's notebook. [c] shows download time of picture which was 12 seconds and [d] is final screen. Comparing preexisted systems, MRSS simplifies step of request and improves convenience. Moreover it shortens download time, too. The downloading time reduction with MRSS can be seen in Fig.9.

(Unit: msec)

Section	Calling Time (Preparation Time)				Server (data location search)	Download Time		Data (Entire download time)		
	Web Browser Open	Instruction (voice/text)	URL Loading	Determine suitability		Content ①	Application ②	①	②	①+②
Normal System	3,000	-	5,000	-	-	2,676	95,626	**10,676**	**103,626**	**114,302**
MRSS	-	2,000	-	591	2,189	2,676	95,626	**7,456**	**95,626**	**103,082**

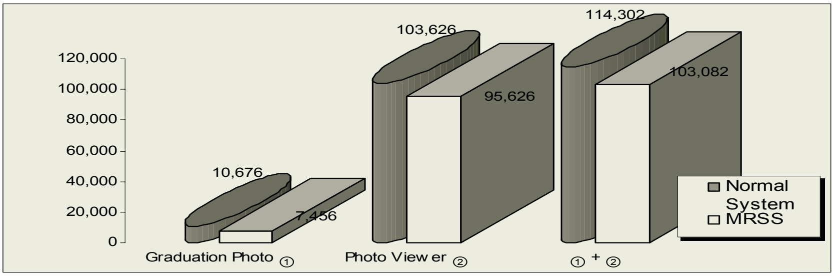


Fig. 9. Download Time Comparison between Normal System and MRSS

5 Conclusion

The expected effects of MRSS are as follows. First, it makes an input processing simpler than before, and enables customers to obtain any files they want with ease. Second, it reinforces connection among mobile resources by enlarging the range of resources. Third, it reutilizes legacy system and finally it is available to apply Ubiquitous Computing.

Further study can be done for more efficient and effective functions of MRSS as follows; first, MRSS shall be changed to the structure that calls dynamic library file beyond general data. Second, MRSS shall be a resource management agent that enables controlling the internal resources of database.

References

1. Y.S. Moon, Introduction to Wireless Internet, Yonhaksa, 2002.
2. Ministry of Information and Communication, http://www.mic.go.kr.
3. C.N. Kim, Study of Next Generation Wireless Internet, Electronic Newspaper, 2002.
4. Mark Weiser, "Hot Topics: Ubiquitous Computing," IEEE Computer, Vol.26, No.10, pp.7172, 1993.
5. Anind K. Dey, "Understanding and Using Context, personal and Ubiquitous Computing," Journal, Vol.5, No.1, pp 47, 2001.
6. Society for the Wireless Internet Business, All about Wireless Internet, Chungang M&B, 2001.
7. E. Kovacs, K. Rohrle and B. Schiemann, "Adaptive Mobile Access to Context-Awareness Services," Proceeding on 1st International Symposium on Agent System and Applications, pp.190201, 1999.
8. S.Y. Jang, U.T, Uoo, "Trend of Sensing technology and Context-Awareness technology for Ubiquitous Computing," Korea Information Processing Society Review, Vol.21, No.5, pp.1828, 2003.
9. S.K. Uoo, Analysis of Korean Phrase by Phrase Relationship, KAIST master degree paper, 1991.

Integrated Management of Multi-level Road Network and Transportation Networks

Jun Feng[1], Yuelong Zhu[1], Naoto Mukai[2], and Toyohide Watanabe[2]

[1] Hohai University, Nanjing, Jiangsu 210024 China
fengjun-cn@vip.sina.com
[2] Nagoya University, Nagoya, Aichi 464-8603 Japan

Abstract. The issue of how to provide location-based service (LBS) for transportation users has attracted many researchers. In a country wide system, there are needs for processing geographic information of road network in different scales and transportation information at the same time. In this paper, we propose an integrated method for representing multi-level of transportation information in addition to multi-scale of road map. Based on the datasets generated by this method, queries in LBS applications can be responded efficiently.

1 Introduction

In a country-wide Intelligent Transportation System (ITS), there is a need for processing the map and transportation information in many levels of details. This is due primarily to the desire of the user to use (/see) relevant information only; too many details may hinder rather than help [1]. Because the generalization of map information cannot be realized automatically [2], various methods have been proposed for maintenance of scaleless maps or multi-scales of maps [1–3]. However, queries in ITS applications often require certain classes of attributes that may not be presented in road maps: e.g., traffic constraints. In the traditional method of representing transportation network, which travel junctions are represented with nodes and links, the traffic constraints are represented by adding new nodes and arcs. The total number of nodes and links in the dataset is multiplied [4, 5], and results in a lower efficiency of processing on the datasets.

In this paper, we extend our representation method for transportation network [6] to represent multi-level transportation network and talk about the search method based on it. This paper is organized as follows. The representation methods of transportation networks and multi-level road networks are presented in Section 2 and Section 3. The integrated management of multi-levels traffic conditions and spatial information about road network is proposed in Section 4. Section 5 talks about spatial search and Section 6 makes a conclusion on our work.

2 Representation of Transportation Information

In this section, we depict a representation method for integrating traffic information and spatial information about road network.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 677–683, 2005.

A road network with nodes and links representing the crosses and road segments can be regarded as a un-directed graph G, $G = (V, L)$, where V is a set of vertices $\{ v_1, v_2, ...v_n\}$, and L is a collection of edges $\{ l_1, l_2, ...l_m\}$. Traffic information on the road network is regarded as a directed graph G', $G' = (V, A)$, where V is a set of vertices $\{ v_1, v_2, ...v_n\}$, and A is a collection of arcs $\{ a_1, a_2, ...a_p\}$.

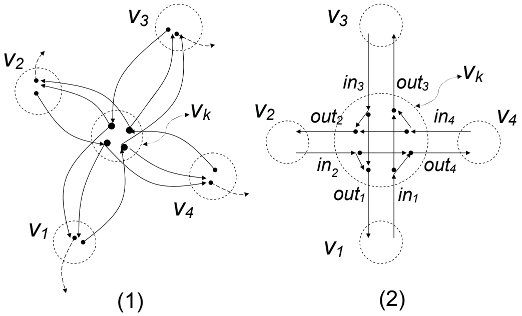


Fig. 1. Cross node with constraint

The typical road junctions with constraints are given in Fig. 1. In Fig. 1(1), road junctions are represented by using [5]'s model. In the graph, each edge depicts a one-way road and each node corresponds to a junction. Two-ways roads can be presented as a pair of edges: one in each direction. This model permits easy modelling of one-way roads and limited access junctions. However, it keeps the topology relations among vertices, and ignores the spatial relations among them. Extra nodes should be added to the graph when there are any access limitations (here, left-turn-forbidden). In other words, one cross node on the road network may be represented with several vertices corresponding to the junctions, and they are independent with each other. Considering the shortcomings of this simple model, we propose a *super-node* representation method for integrating junctions (including traffic cost and traffic constraints) and road network.

A *super-node* can be defined as a node in road network with multiple corresponding junctions: for example, v_k in Fig. 1 (2). The information of the *super-node* contains the following parts (for simplicity of explanation, road junctions in Fig. 1 (2) is used as an example):

1) *Cost-arc*: The arc, which final vertex is v_k, is called in-arcs, denoted as in_i, and similarly the arc, which initial vertex is v_k, is called out-arc, denoted as out_j. The number of these arcs are called as in-degree (e.g. 4) and out-degree (e.g. 4), respectively. Every out_i is defined as a *Cost-arc* consists of the final vertex and the traffic cost for travelling through this arc. *Cost-arc*s of v_k in Fig. 1(2) are $\begin{bmatrix} out_1(v_1, cost_{k1}) \\ out_2(v_2, cost_{k2}) \\ out_3(v_3, cost_{k3}) \\ out_4(v_4, cost_{k4}) \end{bmatrix}$.

2) *Constraint-matrix*: The constraints on the *super-node* can be represented with an $n \times m$ matrix CM:
$$CM(v_k) = \begin{array}{c} \\ in_1 \\ in_2 \\ \vdots \\ in_n \end{array} \begin{array}{c} \begin{array}{cccc} out_1 & out_2 & \dots & out_m \end{array} \\ \begin{pmatrix} C_{11} & C_{12} & \dots & C_{1m} \\ C_{21} & C_{22} & \dots & C_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \dots & C_{nm} \end{pmatrix} \end{array}.$$
And $C_{ij} = \begin{cases} 1 & \text{there is restriction from } in_i \text{ to } out_j; \\ 0 & \text{there is a junction from } in_i \text{ to } out_j. \end{cases}$

More details of this method is given in [6]. This method decreases the redundancies in the database by adopting a complex node representation. It is easy to integrate the traffic information and the basic road network. For the basic road network, the additional information for traffic information is managed on every node. When the number of nodes and traffic arcs is unchanged, the modification of the traffic information does not injure the stability of the spatial index structure for road network. Therefore, a kind of queries in ITS application, which refer to the spatial information, can be solved by taking advantages of the spatial index.

3 Management of Multi-level Road Networks

Contrary to those structures proposed based on the most detailed map, multi-resolution data models provide ways of describing the world at various levels of detail. The importance of such data models in the context of spatial information is widely acknowledged, and there are several studies of their formal foundations [7–9]. Under these models, to realize the sharability among multi-levels or rapid access to multi-levels of maps, map objects are arranged in multi-levels based on a compromise between storage and computation. Although the size of storage, the speed of map zooming or effective spatial process on multi-levels of maps were studied in these works, they share the same complex process to keep the consistency among multi-levels of map information. To solve this problem, an access method, called Multi-levels Object-Relation (MOR-tree) structure, which possesses the ability of the followings, is proposed [3]:

(1) The ability of handling spatial data efficiently: the most of operations on maps are the processing of spatial queries, so the retrieval of data items quickly according to their spatial locations is needed;
(2) The ability to provide integrated access to multi-scale maps: multi-levels of maps could be accessed directly and efficiently;
(3) The ability of arranging the relations among multi-levels of maps: these relations are indispensable for keeping consistency among maps.

MOR-tree is proposed based on M^2 (Multi-level / Multi-theme) map information model [10] for managing multi-levels (of scale) of maps. Under M^2 model, map objects are divided into different levels of scales in regard to the display needs of map [11]. Road objects are assigned to multi-level datasets without repetition: the dataset in the lower level is a supplement of the dataset in the upper level. An example is given in Fig. 2(1), Fig. 2(1)a depicts the road networks in the upper level and Fig. 2(1)b is for that in the lower level. *a1-2* and *b1-3* are road segments in *level 1* (upper level) and *level 2* (lower level), respectively.

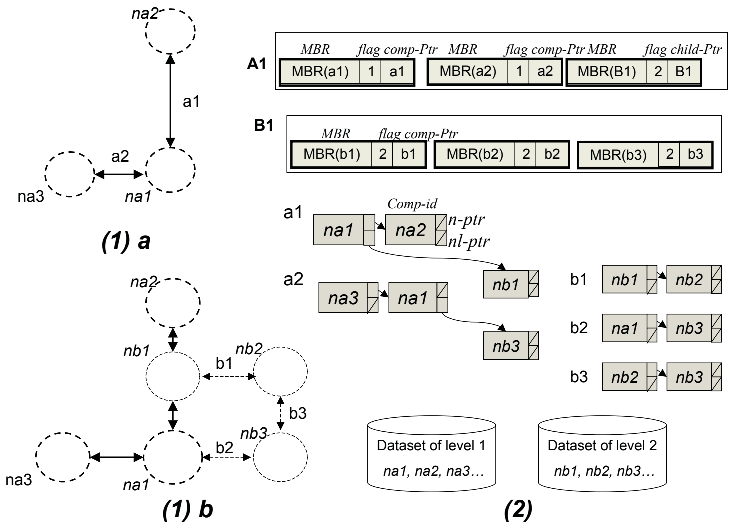


Fig. 2. (1)Multi-level road networks; (2) MOR-tree for multi-level road networks: *A1*: internal nodes of main hierarchy; *B1*: leaf nodes of main hierarchy; *ai*, *bi*: composition hierarchy; *nai*, *nbi*: Composition-entries; *i* (1 or 2): flag of Object-entries or Tree-entries

The MOR-tree for the road networks in Fig. 2(1) is given in Fig. 2(2). There are three kinds of entries with the following forms:

(1) Object-entry has the form (*MBR*, *flag*, *comp-ptr*), where *MBR* is the minimal bounding rectangle of the composition hierarchy; *flag* is a natural number that indicates the importance level; and *comp-ptr* contains a reference to a composition hierarchy;
(2) Tree-entry has the form (*MBR*, *flag*, *child-ptr*), where *child-ptr* contains a reference to a subtree. In this case *MBR* is the minimal bounding rectangle of the whole subtree and *flag* is the smallest importance value of the child-nodes.
(3) Composition-entry has the form (*comp-id*, *n-ptr*, *nl-ptr*), where *comp-id* is the identifier of object's composition; *n-ptr* contains a reference to the next composition of the parent node object: e.g., the object is a road segment, the first composition of it is one of the end points, *n-ptr* points to the next point on the same object; and *nl-ptr* contains a reference to the composition of the parent node object in the lower level: e.g., the intersection between the upper-level road and lower-level road.

However, MOR-tree cannot be applied to transportation information directly, as the transportation information on road network is more complex than the original map. In this paper, an integrated representation method for multi-level transportation network is proposed.

4 Integrated Representation of Multi-level Transportation Networks

Basically, the transportation information is managed on every node of the map, however, the node which connects with road segments in different levels should be processed specially.

Considering the multi-level transportation networks with constraints given in Fig. 3(1), which basic road maps is those in Fig. 2(1): there are road segments: (*na3*, *na1*) and (*na1*, *nb3*), however, because there is constraint on *na1*, there is no transportation arc from *na1* to *nb3*.

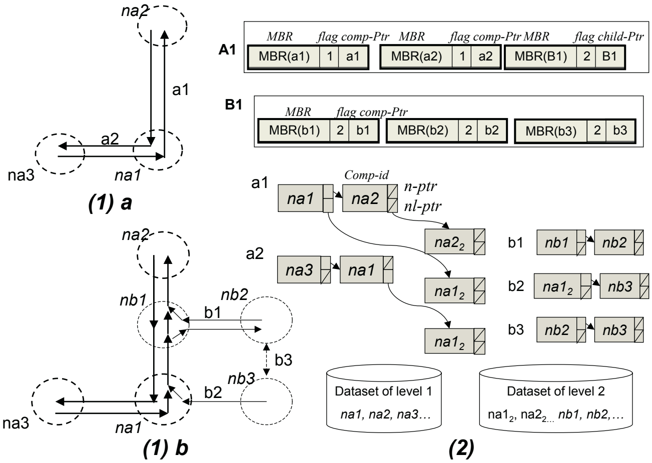


Fig. 3. (1)Multi-level transportation networks; (2) Extended MOR-tree for multi-level transportation networks

Therefore, though in Fig. 2(2), the road segments between nodes of different levels can be recorded in the pointer of *nl-ptr* in composition hierarchy, such pointers cannot assure that there are transportation arcs between these nodes. Moreover, the constraint on the same node maybe different according to different level's usage. Consider that the path search in *level 1* is based on the arcs in *level 1* and the the search in *level 2* is based on a different arc set. For the search in *level 1*, *Cost-arcs* of node *na1* are: $\begin{bmatrix} out_{a1}(na_2, cost_1) \\ out_{a2}(na_3, cost_2) \end{bmatrix}$. while for that in *level 2* the *Cost-arcs* of node *na1* are: $\begin{bmatrix} out_{a11}(nb_1, cost_{11}) \\ out_{a2}(na_3, cost_2) \end{bmatrix}$.

The *Constraint-matrix* of node *na1* in *level 1* and *level 2* are: $CM(na_1)_1 = \begin{array}{c} \\ in_{a1} \\ in_{a2} \end{array}\!\!\begin{array}{c} \begin{array}{cc} out_{a1} & out_{a2} \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}; CM(na_1)_2 = \begin{array}{c} \\ in_{a11} \\ in_{a2} \\ in_{b3} \end{array}\!\!\begin{array}{c} \begin{array}{cc} out_{a11} & out_{a2} \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array}.$

Because 1) the information in the lower level (*level 2*) is not to be used by the operation in *level 1*; and 2) the original MOR-tree is not consistent with the transportation situation, we split one node in the upper level (*level 1*) into two nodes belonging to two levels: one is the original node in the upper level, another is a transportation-node in the lower level, which contains transportation information of the lower level about the original node. In MOR-tree for road network, *nl-ptr* contains a reference to the composition of the parent node object in the lower level, here, we redefine *nl-ptr* as a pointer which contains a reference to a transportation-node of the parent node object in the lower level. The extended composition hierarchy is given in Fig. 3(2). The nodes *na1*, *na2* in *level 1* possess the transportation information referring to *level 2*, so the transportation-nodes for them is created in *level 2*. For the composition hierarchies of *level 1*, *nl-ptr*s refer to the corresponding transportation-nodes in *level 2*. And so the search or operations referring to only the upper level can be done just as using the original MOR-tree, and the operations referring to the two levels can be done by using the information managed in the lower level.

The nodes which only connects with transportation arcs on the same level can be represented with a *super-node* directly.

5 Spatial Search and Path Search

MOR-tree supports not only the spatial search on multi-level maps but also the path search on multi-level transportation information.

Spatial search to multi-level maps can be realized by accessing objects until a specific level via MOR-tree. Because MOR-tree takes the advantages of spatial index structure, spatial queries, such as region query, can be realized by accessing to the internal nodes and composition hierarchies until a specific level of MOR-tree. The zoom in/out operations can be realized by accessing the nodes under a specific internal node of MOR-tree.

Path search on the transportation information on different levels can be realized by using an “inkblot” search method. For example, the path search from A city hall to B city hall in different states, the search starts from A city hall inside city A based on city level and be divided into three parts: two parts of the inter-city search, which are based on the transportation information on the city level inside city A and B, and one part of cross-state search, which is based on the information on one level upper than the city. The changeover points are those nodes on the road network which possess original node and transportation-node on multiple levels.

6 Conclusion

In this paper, we proposed a method for representing multi-level transportation networks and road networks. Our method adopts a spatial index structure for managing map objects in multiple levels and uses an integrated method for representing the traffic constraints. Based on the datasets generated by this method, queries in ITS applications can be responded efficiently in different levels of details. In our future work, the evaluations of the creation, modification and processing of the datasets created by our method will be made.

Acknowledgement

This research is partly supported by the Taiko Foreign Researcher Promotion Foundation (Japan), Science and Technology Promotion Foundation of Hohai University and Technology Foundation of Nanjing, China.

References

1. B. Becker, H.-W. Six, and P. Widmayer. Spatial priority search: an access technique for scaleless maps. *Proc. of ACM SIGMOD'91*, pages 128–137, 1991.
2. P. V. Oosterom. The reactive-tree: a storage structure for a seamless geographic database. *Proc. of Auto-Carto*, 10:393–407, 1991.
3. J. Feng and T. Watanabe. Mor-tree: An access structure for multi-levels of road networks on distributed environment. *Journal of the ISCIE*, 16(12):656–662, 2003.
4. S. Winter. Modeling costs of turns in route planning. *GeoInformatica*, (4):345–361, 2002.
5. J. Fawcett and P. Robinson. Adaptive routing for road traffic. *IEEE Computer Graphics and Applications*, 20(3):46–53, 2000.
6. J. Feng, N. Mukai, and T. Watanabe. Search on transportation network for location-based service. *Proc. of IEA/AIE 2005, LNAI*, to appear, 2005.
7. J. G. Stell. Granulation for graphs. *Lecture Notes in Computer Science, Springer-Verlag*, (1661):417–432, 1999.
8. Y. Leung, K.S. Leung, and J.Z. He. A generic concept-based object-oriented geographical information system. *Int'l J. of Geographical Information Science*, 13(5):475–498, 1999.
9. S. Timpf. Hierarchical structures in map series. *http://www.geoinfo.tuwien.ac.at /publications /formerPersonnel /timpf /diss /table-of-contents.htm*, 1998.
10. J. Feng and T. Watanabe. Effective representation of road network on concept of object orientation. *Trans. IEE of Japan*, 122-C(12):2100–2108, 2002.
11. T. Yamanashi and T. Watanabe. Multi-layers/multi-phases model adaptable to integrated gis. *Proc. of VSMM'00*, pages 668–676, 2000.

Music Plagiarism Detection Using Melody Databases

Jeong-Il Park[1], Sang-Wook Kim[2], and Miyoung Shin[3]

[1] Division of Computer, Information, and Communications Engineering
Kangwon National University, Chuncheon, Korea
inode777@empal.com

[2] College of Information and Communications
Hanyang University, Seoul, Korea
wook@hanyang.ac.kr

[3] School of Electrical Engineering and Computer Science
Kyungpook National University, Daegu, Korea
shinmy@knu.re.kr

Abstract. This paper addresses the development of a system that detects plagiarism based on similar melody searching. Similar melody searching is to find the melodies similar to a given query melody from a music database. For this purpose, we propose a novel similarity model that supports alignment as well as shifting. Also, we suggest a method for indexing the features extracted from every melody, and a method for processing plagiarism detection by using the index. With our plagiarism detection system, composers can easily search for the melodies similar to their ones from music databases. Through performance evaluation via a series of experiments, we show the effectiveness of our approach. The results reveal that our approach outperforms the sequential-scan-based one in speed up to around 31 times.

1 Introduction

Recently, the use of various types of multimedia data such as images, videos, and audios has shown its explosive growth, so the *content-based search* became of great importance[12][7][13][5][4][3]. For the successful content-based search, an indexing scheme and a query processing scheme are the key issues to be considered. Despite of its great advance, audio search has been less investigated than either image or video searches [8].

In this paper, we address the development of a system that detects plagiarism based on the *similar melody searching*, which is an operation that finds the melodies similar to a given query melody from a music database. Specifically, the plagiarism detection system is to examine if there exist such melodies in music databases as being similar to any melodies of a composer's interest. Without realizing it, any composers may be involved in diverse plagiarism disputes. The main purpose of this research is to help composers to avoid unnecessary plagiarism disputes by using the plagiarism detection system in advance.

Unlike such previous systems as [6][8][9][10], our plagiarism detection system has its unique characteristics as follows:

- A novel similarity model: it solves the problem of misjudgment by supporting alignment as well as shifting in the similarity model.
- Multidimensional indexing: it makes a basic framework for fast searching by employing a multidimensional index built on melody features.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 684–693, 2005.

- A three-step query processing: it provides fast search ability by taking a three-step query processing method, which consists of index searching, window stitching, and post-processing.

The results of performance evaluation show that our approach outperforms the sequential-scan-based one in the speed of searching up to around 31 times.

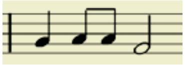
This paper is organized as follows. Section 2 describes a novel similarity model for computing the similarity between two different melodies. Section 3 discusses an indexing method for efficient similar melody searching from melody databases. Section 4 presents a three-step query processing method which employs the proposed indexing method. To show its superiority, we evaluate the performance of our system in Section 5. Finally, Section 6 summarizes and concludes the paper.

2 Similarity Model

This section describes a similarity model to compute the similarity between two different melodies.

2.1 Basic Model

A melody of a music is defined as a list of snatches, i.e. as a sequence $S = <(s_i, sL_i)>$ $(0<=i<n)$, called *a melody sequence*. Here, s_i denotes the i^{th} tone of a melody, and sL_i denotes the length of the i^{th} tone of a melody. Also, the number of tones, n, is called the length of a melody sequence. For example, the following melody with a C major key in four-quarters time is described as a melody sequence of length 4, i.e. <(sol, 1), (la, 1/2), (la, 1/2), (fa, 2)>[1].



For comparison of any two melody sequences $S = <(s_i, sL_i)>(0<=i<n)$ and $Q = <(q_j, qL_j)>(0<=j<m)$, we make the following assumptions:

- *Assumption 1:* S and Q should have the same meter. In other words, two melody sequences of different meters are excluded for similarity searching.
- *Assumption 2:* It always holds that n=m and $sL_i = qL_j$ if i=j, for all (i, j)'s. Thus, two melody sequences to be compared should be of the same length, and the lengths of their corresponding tones should be also the same.

For any two melody sequences S and Q satisfying these assumptions, the similarity between S and Q is computed based on Definition 1 described below.

Definition 1: For two melody sequences $S = <(s_i, sL_i)>(0<=i<n)$ and $Q = <(q_j, qL_j)>(0<=j<n)$, they are defined as being similar to each other if and only if the following condition, as well as Assumptions 1 and 2, are satisfied.

$$L_\infty(S, Q) = L_\infty(<s_0, s_1, ..., s_{n-1}>, <q_0, q_1, ..., q_{n-1}>) < \varepsilon$$

[1] For illustration, in this paper, a tone and its length are described as a symbol and a fraction, respectively. However, in practice, when being stored into standard MIDI file [11], both of them are described as integers, respectively.

Here, ε is a tolerance specified by users, and L_∞ is a function that returns the maximum difference between the corresponding tones of two sequences [1].

That is, two melodies are considered as being similar only if the maximum difference between the corresponding tones of two melodies is less than or equal to ε. This implies that any tone in a melody S should always have its corresponding tone in a melody Q within ε.

2.2 Alignment

As mentioned in Section 2.1, in our basic similarity model, the similarity between two melody sequences can be computed only when they are of the same length and their corresponding tones are also of the same length. However, such assumptions are often too restrictive in practice to include "real" similar melodies for comparisons. For example, consider the following case shown in Fig. 1. Here, two melodies are almost identical, but Assumption 2 makes these excluded from the targets for comparisons.

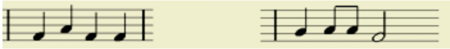


Fig. 1. Two melodies of different lengths

To handle such a problem, we employ a new operation called *alignment*. The alignment is to force two melody sequences and their corresponding tones to be of the same lengths, respectively, so that they can be eventually compared. Yet, note that this alignment can be applied only when two melodies are of the same meters and of the same number of measures.

The alignment is done based on the *split* of a tone. The split is to divide a tone into many shorter tones in case of having the tones of different lengths to be compared. Fig. 2 shows an example of alignment where two different melodies of different lengths are split. By the alignment, therefore, it becomes possible to apply similarity searching for any melodies of the same meters and the same measures[2].

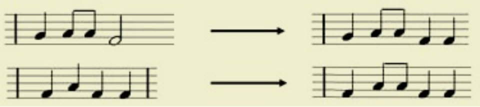


Fig. 2. An example of alignment

Hereafter, further discussions are made based on the premise that the alignment has already been applied for any two melodies S and Q to be compared so that the two melodies should be of the same length and also their corresponding tones should be of the same length.

[2] Along with a split operation, some previous studies also introduced a merge operation that is opposite to a split [9]. However, the split-based alignment alone is also able to handle all such cases.

2.3 Shifting

In our basic similarity model presented in Section 2.1, the similarity is measured only based on the absolute pitch of a tone. So, it often says two *really* similar melodies of different tones to be dissimilar. For example, suppose that two melodies A and B are completely equal to each other in the changing pattern and the length of their corresponding tones, except in the absolute pitch of tone. In such a case, we can tell that the two melodies are very similar, but our basic similarity model does not say so, i.e. it tells these melodies are not similar. To deal with this problem, we use a *shifting* operation here. The shifting is to transform a melody sequence S (=<(s_i, sL_i)>, (0<=i<n)) into S' by using the following equation:

$$\mathrm{SHIFT}(S) = S' = <(s_i', sL_i)>\ (0<=i<n), \text{ where } s_i' = s_i - (max_s + min_s)/2$$

Here max_s and min_s denote the maximum and the minimum pitches of the tones, respectively, in a melody sequence S.

By applying a shifting operation, for any two melodies which have similar changing patterns but different absolute pitches of their corresponding tones, our similarity model is able to correctly judge about the similarity between them. Thus, we redefine the criteria about the similarity between two melody sequences S and Q as in Definition 2.

Definition 2: Two melody sequences S = <(s_i, sL_i)>(0<=i<n) and Q = <(q_j, qL_j)>(0<=j<n) are defined to be similar if they satisfies the following condition

$$L_\infty(\mathrm{SHIFT}(S), \mathrm{SHIFT}(Q)) = L_\infty\ (<s_0', s_1', ..., s_{n-1}'>, <q_0', q_1', ..., q_{n-1}'>) < \varepsilon$$

Here, $s_i' = s_i - (max_s + min_s)/2$ and $q_i' = q_i - (max_q + min_q)/2$. Also, ε is a tolerance specified by users, and L_∞ is a function that returns the maximum difference between the corresponding tones in the two melody sequences.

By employing the similarity model which allows a shifting operation, the similar melody searching can be done with a focus on the changing pattern of the tones, rather than on the pitch itself of the tones, in each melody sequence.

3 Indexing Method

This section discusses an indexing method to effectively search for a set of melody sequences similar to a given query melody.

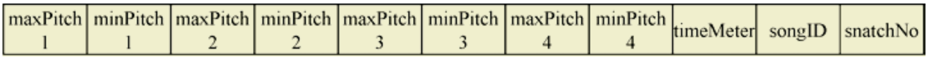
3.1 Feature Extraction

To begin with, we define a *window* as a series of k measures in a melody sequence. The window is considered as an indexing unit in this study. Thus, once some features are extracted from each measure, the indexing is made for each of these features in a series of k measures. The features taken from each measure of a melody are timeMeter, maxPitch, and minPitch. Here timeMeter denotes the meter of a measure, and maxPitch and minPitch denote the maximum and the minimum pitches of the tones in a measure, respectively.

3.2 Index Structure

Since we consider multiple features for indexing as described earlier, we use the R*-tree [2] as an indexing structure. As a multidimensional index structure, the R*-tree is often employed for an effective indexing of multiple features.

For illustration, Fig. 3 presents an entry structure of leaf node in our R*-tree under the assumption that a window has 4 measures.

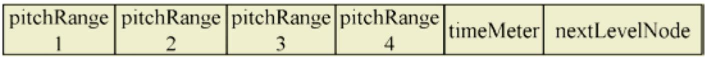


maxPitch 1	minPitch 1	maxPitch 2	minPitch 2	maxPitch 3	minPitch 3	maxPitch 4	minPitch 4	timeMeter	songID	snatchNo

Fig. 3. An entry structure of a leaf node

That is, the entry structure includes (1) a pair of <maxPitch, minPitch> for 4 measures in a window, (2) a meter and (3) an identifier of the song which includes the window, and (4) the starting position of the window within the song. songID is an identifier for indicating a song within the entire database, and snatchNo is an identifier for indicating the position of a measure within the corresponding song.

Among the attributes shown in Fig. 3, the organizing attributes [14] for the R*-tree are determined as follows. First, songID and snatchNo are excluded from the organizing attributes since these are not used to specify a search condition. The values of maxPitch i and minPitch i are replaced with a single value of pitchRange i(= maxPitch i - minPitch i). By doing this, the number of organizing attributes can be reduced from k*2 to k, so we take k pitchRange i's as organizing attributes instead of k maxPitch i's and k minPitch i's. Also, timeMeter is taken as the $(k+1)^{th}$ organizing attribute. Consequently, a window can be projected into a point on the (k+1) dimensional space, and we call it a *data window point*. Fig. 4 presents an entry structure of a non-leaf node in the R*-tree. The last field, nextLevelNode, represents a pointer to the next level node in the R*-tree.



pitchRange 1	pitchRange 2	pitchRange 3	pitchRange 4	timeMeter	nextLevelNode

Fig. 4. An entry structure of a non-leaf node

The R*-tree stores and manages numerous data window points existing on the (k+1) dimensional space. From these data window points, the R*-tree search returns candidate data window points that are most likely to be included in the final query result. On the other hand, the features of maxPitch i and minPitch i shown in Fig. 3 are not directly used for the R*-tree search. Instead, they are used to reduce the number of candidate window points in the window stitching step, which will be described in Section 4.

3.3 Index Construction

The procedure of building the R*-tree for all the songs stored in an entire database is shown in Algorithm 1. Here, the sliding windows indicate the windows extracted from all the possible measures of a song. On the other hand, the disjoint windows mentioned in Section 4 indicate the windows extracted from the song without overlapping.

```
FOR each song accessed from a music database,
1. Extract sliding windows, each of which consists of k measures, from the song;
2. FOR each sliding window,
   2.1. Obtain Fsw, the features mentioned in Figures 3 and 4, from the sliding window;
   2.2. Insert Fsw into the R*-tree;
```

Algorithm 1. The R*-tree construction

4 Query Processing

This section addresses a query processing method for the effective handling of similar melody searching based on the index structure proposed in Section 3. The melody given in a query for similar melody searching is defined as a *query melody*. Query processing proceeds in the order of followings: feature extraction, R*-tree searching, window stitching, and post-processing. Each of these steps is detailed in this section.

4.1 Feature Extraction

For feature extraction, we first extract *disjoint windows (dw)* consisting of k measures from a query melody. Recall that disjoint windows indicate the windows extracted without overlapping. The maxPitch and the minPitch within a series of k measures for each dw, and the timeMeter of the dw are taken as selected features for window stitching. Also, by obtaining pitchRange i from maxPitch i and minPitch i, we represent each dw as a point on the (k+1) dimensional space, called a *query window point*.

For each query window point, we obtain the k-dimensional square whose side has a distance ε from the query window point. However, note that timeMeter is not considered for this ε-extension. This is because we need to find such songs that exactly match in their meters but have a tolerance of ε in their pitchRange. Thus, the k-dimensional square with ε-extension is called a *query square*.

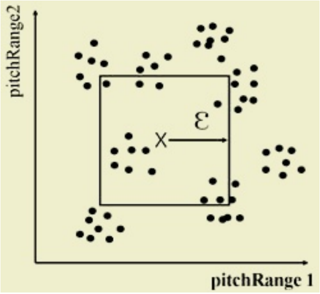


Fig. 5. A query square along with data window points

4.2 Index Searching

Index searching is the first filtering procedure for finding such candidate melodies that have a high possibility of being similar to a query melody. By traversing the R*-tree index, we find data window points included in each query square, which are called *candidate data window points*. Fig. 5 shows a distribution of data window points and a query window point in a multidimensional space. For illustration, we assume them in a two dimensional space. Here a point marked by X indicates a query window point while the other points indicate data window points. Also, the square denotes a query square, and the points within the square denote candidate data window points returned from index searching.

4.3 Window Stitching

Window stitching is the second filtering procedure for identifying candidate melodies which are highly likely to be similar to a query melody. This procedure eliminates, from candidate data window points returned from index searching, the points corresponding to data melodies which are very unlikely to be similar to a query melody. The basic idea of window stitching is to select the melodies that include a series of data windows matched with those in a query melody, from the candidate windows that are found to be similar to each query window.

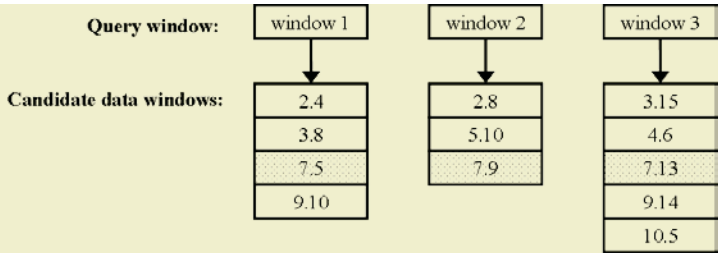


Fig. 6. An example of window stitching

Fig. 6 shows an example of window stitching. Here the three boxes on the top represent three disjoint windows extracted from a query melody. Also, each set of boxes at the bottom represent candidate data windows found by index searching to be similar to the corresponding query window. Each i-j in the box of a candidate data window means that the corresponding candidate data window is a window starting at the j^{th} measure of the i^{th} song. Here each window is assumed to consist of four measures. The only melody that includes a series of candidate data windows similar to all the three query windows is the one starting at the fifth measure of the seventh song (see dotted boxes for reference). Thus, except this one, all the remaining candidate data windows are eliminated soon after the window stitching step since they do not meet the above condition. As a result, the number of candidate melodies is significantly reduced, which is the number of being actually accessed from disk and computed for their similarity to a query melody in post-processing step.

4.4 Post-processing

Post-processing is the final procedure for determining the melodies actually similar to a query melody. This procedure lastly examines if each of the candidate melodies returned from the window stitching step is really similar to a query melody by accessing it from disk. Thus, the melodies whose difference to a query melody exceeds ε are excluded in the final result.

5 Performance Evaluation

In this section, the performance of the proposed method is evaluated. Section 5.1 describes our experiment environment and Section 5.2 presents experiment results.

5.1 Experiment Environment

For experiments, we used a database that stores a total of 500 music files in MIDI format 1 [11]. The query melodies were generated by taking such melodies that starts at a random position of the song arbitrarily chosen from the database, and a tolerance ε were taken as 0, 1, and 2. An experimental result is evaluated with the averaged response times over 10 query melodies of the same length. The query melodies of length 4 and 8 were used, and the window size of 2 was used. For comparative assessment, we considered the same similarity model with a sequential-scan-based method[3]. As a hardware platform for our experiments, we used a 2GHz pentium PC with 512 bytes. Also, Linux, mySQL, and C++ were used as a software platform.

5.2 Experiment Results

In experiment 1, we measured a total processing time of similar melody searching for a query melody consisting of 4 measures, which are shown in Fig. 7. Here the horizontal axis indicates the number of songs stored in a database and the vertical axis indicates the total processing time spent for similar melody searching. IB_SMS (index-based similar melody searching) exhibits the processing time of similar melody searching based on our proposed method while Seq exhibits the processing time of similar melody searching based on the sequential scan. Also, the number in parentheses exhibits a tolerance ε used in a query.

Our experimental results show that the processing time in Seq drastically increases with an increase of the number of songs in a database. This is because the CPU processing time for similarity comparisons and the disk access time are required for all the songs in the database. On the other hand, the processing time of IB_SMS is almost unchanging with the number of songs in a database. This is because the proposed method considers as the candidates only the melodies that have high possibility of being similar to a query melody through the filtering procedures such as index searching and window stitching. Due to our proposed indexing structure and query processing method, we could obtain about 13 to 28 times better performance in similar melody searching, depending on the value of a tolerance.

In experiment 2, we measured the processing time of similar melody searching for a query melody with 8 measures. Fig. 8 shows its results. Overall, the performance tend to be very similar to that of experiment 1, except that, for both IB_SMS and Seq, more processing time was taken than for a query melody with 4 measures. In the case of Seq, this is caused by the fact that the time for computing the similarity between melodies is proportional to the length of a melody. In the case of IB_SMS, it can be interpreted as follows. First, the time for both index searching and window stitching

[3] Direct comparisons with previous methods have not been made because our approach employs a different similarity model from the previous methods.

increases because the number of query windows doubles. Also, the time for computing the similarity between melodies is proportional to their lengths. Accordingly, these factors eventually lead to a considerable growth of the total processing time. Compared with the sequential-scan-based method, our approach shows approximately 18 to 31 times better performance, depending on the value of a tolerance. Consequently, these results show a superiority of our indexing and query processing methods.

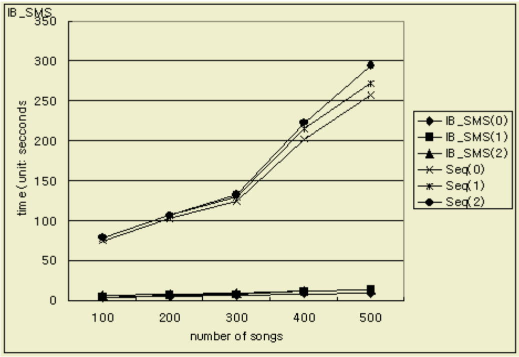


Fig. 7. Results with a query melody having 4 measures

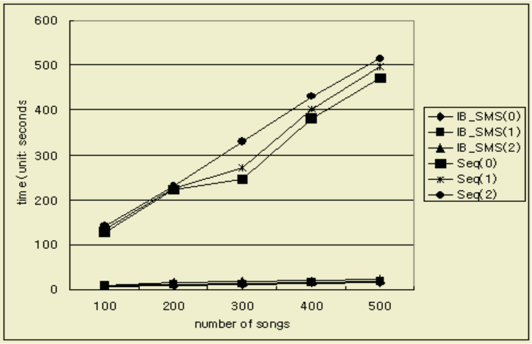


Fig. 8. Results with a query melody having 8 measures

6 Conclusions

Similar melody searching is an operation that finds such melodies similar to a given query melody from a music database. In this paper, we have discussed developing a plagiarism detection system based on similar melody searching. We have first proposed a novel similarity model that supports alignment as well as shifting. Also, we have not only suggested a method of indexing the features extracted from a melody by using a multidimensional index, the R*-tree, but also proposed a three-step query processing method using such an index which consists of index searching, window stitching, and post-processing steps. By our plagiarism detection system, users can effectively search for the melodies from a database which are similar to a melody of their interest.

In order to investigate the superiority of our approach, we evaluated the performance via a series of experiments. The results reveal that the proposed method outperforms the sequential-scan-based one in its speed, up to around 31 times depending on a tolerance used.

Acknowledgement

This work has been supported by the IT Research Center via Kangwon National University, Korea Research Foundation with Grant KRF-2003-041-D00486, and the IT Research Center via Cheju National University. Sang-Wook Kim would like to thank Jung-Hee Seo, Suk-Yeon Hwang, Grace(Joo-Young) Kim, and Joo-Sung Kim for their encouragement and support.

References

1. R. Agrawal et al.: Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. Proc. Int'l. Conf. on Very Large Data Bases, VLDB (1995) 490-501
2. N. Beckmann et al.: The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles. Proc. Int'l. Conf. on Management of Data, ACM SIGMOD (1990) 322-331
3. G. H. Cha, C. W. Chung: A New Indexing Scheme for Content-Based Image Retrieval. Multimedia Tool and Application 6(3) (1998) 263-288
4. Y. F. Day et al.: Object-Oriented Conceptual Modeling of Video Data. Proc. Int'l. Conf. on Data Engineering, IEEE (1995) 401-408
5. M Flickner et al.: Query by Image and Video Content: The Qbic System. IEEE Computer 28(9) (1995) 23-32.
6. A. Ghias et al.: Query by Humming: Musical Information Retrieval in an Audio Database. ACM Multimedia (1995) 231-236
7. R. Hjelsvold and R. Midtsraum: Modeling and Querying Video Data. Proc. Int'l. Conf. on Very Large Data Bases, VLDB (1994) 686-694
8. J. L. Hsu et al.: Efficient Repeating Pattern Finding in Music Databases. Proc. Int'l. Conf. on Information and Knowledge Management, CIKM (1998) 281-288
9. S. Y. Kim and Y. S. Kim: An Indexing and Retrieval Mechanism Using Representative Melodies for Music Databases. Proc. Int'l. Conf. on Information Society in the 21st Century, (2000)
10. C. C. Liu, P. J. Tsai: Content-Based Retrieval of MP3 Music Objects. Proc. Int'l. Conf. on Information and Knowledge Management, ACM CIKM (2001) 506-511
11. Standard MIDI File Format, Standard MiDI Files 1.0 (1988)
12. E. Oomoto and K. Tanaka: OVID: Design and Implementation of a Video-Object Database System. IEEE Trans. on Knowledge and Data Engineering (1993) 629-643
13. S. W. Smoliar and H. J. Zhang: Content-Based Video Indexing and Retrieval. IEEE Multimedia (1994) 62-71
14. K. Y. Whang, S. W. Kim, and G. Wiederhold: Dynamic Maintenance of Data Distribution for Selectivity Estimation. The VLDB Journal 3(1) (1994) 29-51

News Video Retrieval Using Automatic Indexing of Korean Closed-Caption

Jungwon Cho[1], Seungdo Jeong[2], and Byunguk Choi[3]

[1] Department of Computer Education, College of Education,
Cheju National University, 66 Jejudaehakno, Jeju-si, Jeju-do, 690-756 Korea
jwcho@cheju.ac.kr

[2] Multimedia Laboratory, Department of Electrical and Computer Engineering,
Hanyang University, 17 Haengdang-dong, Sungdong-gu, Seoul, 133-791 Korea
kain@mlab.hanyang.ac.kr

[3] Division of Information and Communications, Hanyang University,
17 Haengdang-dong, Sungdong-gu, Seoul, 133-791 Korea
buchoi@mlab.hanyang.ac.kr

Abstract. Knowledge-based video retrieval is able to provide the retrieval result that corresponds with conceptual demand of user because of performing automatic indexing with audio-visual data, closed-caption, and so on. In this paper, we present the automatic indexing method of Korean closed-caption for knowledge-based video retrieval and the retrieval scheme using the indexed database. In the experiment, we have applied the proposed method to news video with the closed-caption generated by Korean stenographic system, and have empirically confirmed that the proposed method could provide the retrieval result that corresponds with more meaningful conceptual demand of user.

1 Introduction

Digital multimedia data still have been grown. The large amount of data needs efficient management and utilization. The information retrieval system is defined as a system that retrieves data user wants to find by extracting features from the collected data and storing the features. Main interest of this system is to retrieve data quickly that correspond with the user's demand. The main subjects of information retrieval have been changed from text to image and video because of popularization of internet and infrastructure with superhighway network.

Information retrieval technique has been changed from keyword-based text retrieval, which extracts keyword by using text processing, to content-based image retrieval [1], which extracts low-level features such as color, shape, and texture of image. It is hard to provide the retrieval result that corresponds with conceptual demand of user by only the content-based retrieval using low-level features. A knowledge-based retrieval [2], full-contents-based retrieval, uses all contents included in video, such as moving picture, audio, closed-caption, and so on. The knowledge-based retrieval method is able to obtain the retrieval result corresponded with conceptual demand of user.

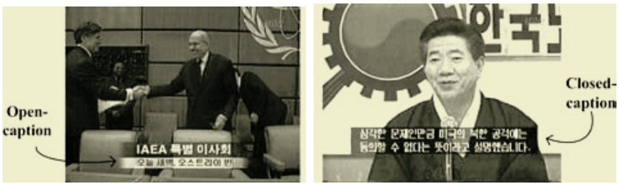
R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 694–703, 2005.

In this paper, we present the automatic indexing method of Korean closed-caption for knowledge-based video retrieval and the retrieval scheme using the indexed database. To use the close-caption, the time-alignment method synchronizing the closed-caption with voice in video is essential. The time-alignment of closed-caption is directly performed by assigning the received time as a unit of word in the caption decoder. The alignment method is able to minimize the false synchronization error. Basic search, which performs the video retrieval on keyword included in the closed-caption, is able to provide the retrieval result by using only the time-alignment of closed-caption. In order to obtain the retrieval result that corresponds more with conceptual demand of user, the morphological analysis and the disambiguation of the result need to be performed to the closed-caption. Then processing of unknown-word, complex-noun, and stopword are performed to select the final index term. The stopword is useless as an index term. Advanced search provides more precise result by using the indexed database, and also supports the content-based image retrieval using the color-structure descriptor of MPEG-7.

2 Caption: Useful Data for Video Retrieval

Because the closed-caption includes not only voice of speakers in video but also distinction of speaker, spatial position of speaker in visual screen, and description of circumstance status, video retrieval system using the closed-caption is effective.

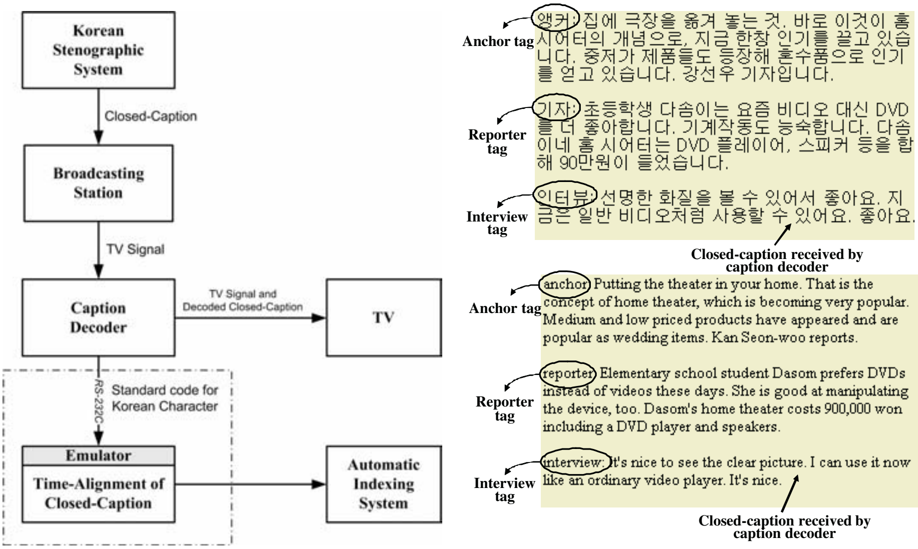
Captions in video are divided into the open-caption and the closed-caption. The closed-caption needs to be decoded by the caption decoder, whereas the open-caption is displayed on screen without the decoder. Fig. 1 shows the open-caption and the closed-caption on TV screen.


Open-caption
IAEA 특별 이사회
오늘 새벽 오스트리아 빈
Closed-caption
심각한 문제인만큼 미국의 북한 공격에는
동의할 수 없다는 뜻이라고 설명했습니다.

(a) open-caption (b) closed-caption

Fig. 1. Captions in video

The closed-caption service has been launched for the hearing impaired-people. Because the close-caption reflects voice of speaker in video, this caption is the best appropriate data to present the meaning of video. Note that the closed-caption, consisted of the standard code for character without any other recognition, is obtained by the caption decoder as shown in Fig. 2. This closed-caption is directly used for basic search or inputted to the morphological analysis system for advanced search after synchronized with the voice in video. In the case of news video used as test-set in this paper, as shown in Fig. 2, it can effectively be used for segmentation of video because the closed-caption is tagged with anchor, reporter, and interview. In the case of closed-caption in movie or drama, also, it can effectively be used for searching because the caption includes not only voice of speaker but also description about visual and sound data.


Korean Stenographic System
Closed-Caption
Broadcasting Station
TV Signal
Caption Decoder
TV Signal and Decoded Closed-Caption
TV
RS-232C
Standard code for Korean Character
Emulator
Time-Alignment of Closed-Caption
Automatic Indexing System
Anchor tag
Reporter tag
Interview tag
Closed-caption received by caption decoder
Anchor tag
Reporter tag
Interview tag
Closed-caption received by caption decoder

앵커: 집에 극장을 옮겨 놓는 것. 바로 이것이 홈시어터의 개념으로, 지금 한창 인기를 끌고 있습니다. 중저가 제품들도 등장해 혼수품으로 인기를 얻고 있습니다. 강선우 기자입니다.

기자: 초등학생 다솜이는 요즘 비디오 대신 DVD를 더 좋아합니다. 기계작동도 능숙합니다. 다솜이네 홈 시어터는 DVD 플레이어, 스피커 등을 합해 90만원이 들었습니다.

인터뷰: 선명한 화질을 볼 수 있어서 좋아요. 지금은 일반 비디오처럼 사용할 수 있어요. 좋아요.

anchor Putting the theater in your home. That is the concept of home theater, which is becoming very popular. Medium and low priced products have appeared and are popular as wedding items. Kan Seon-woo reports.

reporter Elementary school student Dasom prefers DVDs instead of videos these days. She is good at manipulating the device, too. Dasom's home theater costs 900,000 won including a DVD player and speakers.

interview: It's nice to see the clear picture. I can use it now like an ordinary video player. It's nice.

Fig. 2. Schematic diagram of the closed-caption extraction and tags in the caption

3 Time-Alignment of Closed-Caption

The time-alignment, synchronization between closed-caption and voice in video, is essential for information retrieval using the closed-caption. In the *Informedia* project [2, 3], using the closed-caption, the time-alignment is performed by voice recognition. Although this method successfully performs time-alignment of the closed-caption, processing time is longer because of voice recognition, and there is drawback affected by environment in which voice is recorded. Especially, there are many noises in reporter's voice that is recorded at the outside of studio and the voice includes other's voice as background. Therefore, it is difficult to resolve false synchronization error with the time-alignment using voice recognition. In this paper, we analyze the stenographic system generating Korean closed-caption and propose the more accuracy and faster the time-alignment of closed-caption.

3.1 Overview of Korean Stenographic System

Fig. 3 illustrates a schematic diagram of Korean stenographic system. A stenographic system consists of two teams, and each team has two stenographers.

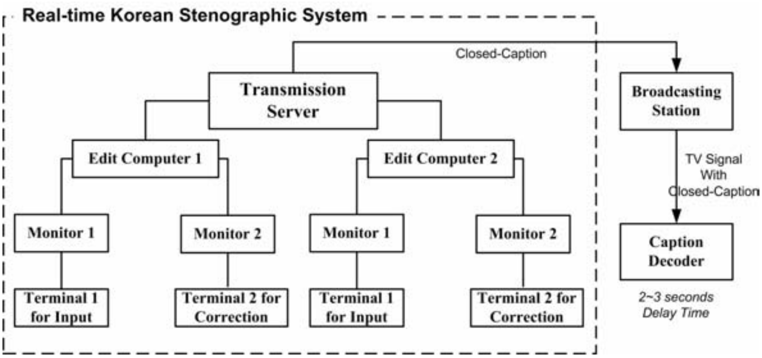


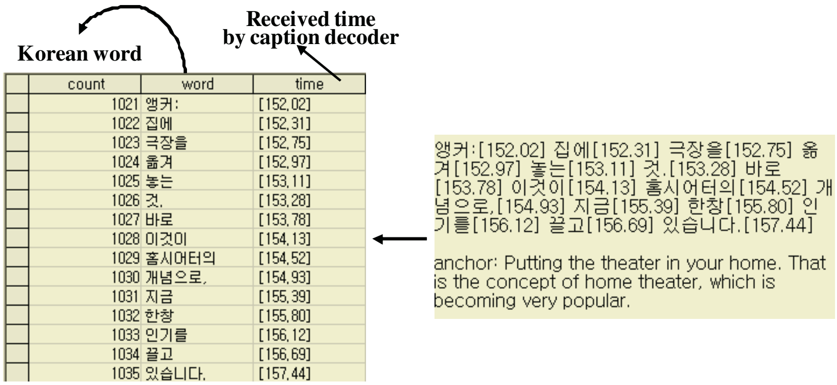
Fig. 3. Schematic diagram of Korean stenographic system

For more accurate input, each team alternatively performs input and correction at the interval of 3 seconds in order to keep stenographers concentrating on the work. In a team, one inputs text, and the other edits inputted text for correction. The completed text is transmitted from a server to a broadcasting station. As shown in Fig. 3, the Korean stenographic system needs three computers and four stenographic terminals. One of three computers is for transmission to broadcasting station, and the others for the input and correction. Note that the Korean stenographic system, the real-time system inserting the closed-caption on live broadcasting, is estimated that the accuracy is about 99.5% to 99.9%, and the delay time is 2∼3 seconds. After each 3 seconds amount of closed-caption generated by stenographers in the system is transmitted to broadcasting station, audience can see the caption on TV screen 2∼3 second later at home.

3.2 Time-Alignment

As mentioned above section, the closed-caption generated by the real-time stenograph system has 2∼3 seconds of delay time, compared with voice in video. In Korean stenograph system, namely, there is a regular interval between time stenographers input and correct on and time stenographers transmit on. Therefore, we are able to obtain the closed-caption synchronized with voice of video by assigning the received time to the closed-caption and subtracting an amount of the delay time from the received time. In the case of a general filmed broadcasting, because the closed-caption is produced in advance, we are able to synchronize the closed-caption with the received time without additional processing to delay time. This time-assignment method, processing in real-time, not only does not need other processes such as voice recognition, but also obtains the closed-caption aligned to synchronized time with voice accurately and quickly.

The time-alignment of closed-caption is performed to the unit of a word of Korean. A consideration, when the additional processing is performed for synchronizing, is that the time-alignment must not be assigned later than the voice in video. The reason is that we loss the start point where voice-visual data must be played at. Therefore, the additional processing must be performed by maximum delay time, 3 seconds. Fig. 4 shows data stored in database after performing the time-alignment as a unit of a word. The recorded unit of received time is a 1/100 second.



	count	word	time
	1021	앵커:	[152,02]
	1022	집에	[152,31]
	1023	극장을	[152,75]
	1024	옮겨	[152,97]
	1025	놓는	[153,11]
	1026	것.	[153,28]
	1027	바로	[153,78]
	1028	이것이	[154,13]
	1029	홈시어터의	[154,52]
	1030	개념으로,	[154,93]
	1031	지금	[155,39]
	1032	한창	[155,80]
	1033	인기를	[156,12]
	1034	끌고	[156,69]
	1035	있습니다.	[157,44]

앵커:[152.02] 집에[152.31] 극장을[152.75] 옮겨[152.97] 놓는[153.11] 것.[153.28] 바로[153.78] 이것이[154.13] 홈시어터의[154.52] 개념으로,[154.93] 지금[155.39] 한창[155.80] 인기를[156.12] 끌고[156.69] 있습니다.[157.44]

anchor: Putting the theater in your home. That is the concept of home theater, which is becoming very popular.

Fig. 4. Closed-caption in database after time-alignment

Using this closed-caption synchronized with voice in video, we are able to effectively retrieve the video with basic search.

4 Automatic Indexing Scheme

When the time-alignment of closed-caption is completed, basic search is able to be performed by string matching of the closed-caption with query. However, after video segmentation, the language processing to the closed-caption is needed in order to retrieve the high priority scenes.

In this paper, we use news video including the closed-caption generated by the real-time Korean stenographic system. Because the video segmentation is performed logically and voice data represented by the closed-caption have a large amount of information in news video, it is proper way to confirm the performance of information retrieval system using the closed-caption. Fig. 5 illustrates a schematic diagram from video to the proposed indexing system.

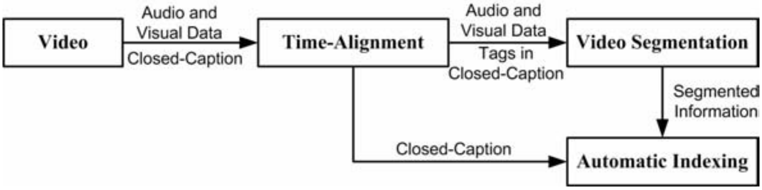


Fig. 5. Schematic diagram from video to the automatic indexing system

4.1 Video Segmentation Using Closed-Caption

In general, news video consists of articles, which have an anchor part, a reporter part, and an interview part. Fig. 6 shows a general structure of news video.

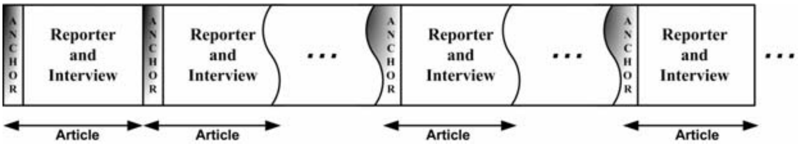


Fig. 6. Structure of news video

In order to use structural characteristic in news video as shown in Fig. 6, previous news retrieval system extracts frames having shape of anchor. The anchor frame is the beginning of article in news video. Using the anchor frame, the previous system successfully segments news video into articles [4]. However, if the frame similar to the anchor frame exists and the false anchor frame is extracted from news video, the previous system has low performance of segmentation of news video. Besides, because the system processes all frames in video, a large amount of processing time is required.

In order to discriminate who speaker is, the closed-caption includes tags that does not exist in voice data. In the case of news video in Korea, the closed-caption has '앵커: ' (anchor), '기자: ' (reporter), and '인터뷰: ' (interview) tags. The proposed method segments news video by using the anchor tag, whereas the previous method segments news video by using the anchor frame. Note that the proposed method is able to resolve the segmentation error resulted from false extraction of anchor frame.

4.2 Keyword Extraction

The keyword is extracted from the closed-caption of articles, after news video is segmented into articles. First of all, in order to extract the keyword, the closed-caption is inputted to the Korean morphological analysis system as the unit of word.

Fig. 7 illustrates a schematic diagram of keyword extraction system. Because the results of morphological analysis still have ambiguity, secondly, disambiguation rule in [5] is applied. If stopword list is applied to the nouns extracted from the disambiguation result, the nouns which are not stopword can be used as keyword. However, for more precise keyword extraction, we added processes required in Korean automatic indexing. In the results from the disambiguation process, unknown-words, which are not in electronic dictionary, are estimated. Then parsing and combination of complex-noun are performed, candidate nouns are selected, and stopwords of candidate nouns are removed. As a result, we select the nouns of the final result. Note that the nouns are keywords for retrieval.

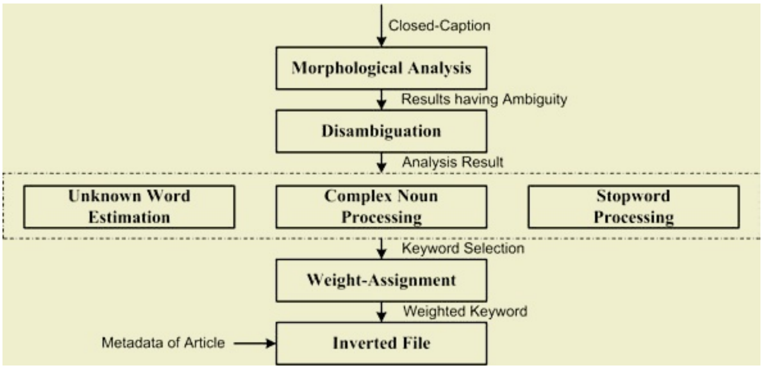


Fig. 7. Schematic diagram of the keyword extraction system

Eq. (1) is the modified equation of the inverted document frequency. The weight of selected keyword is computed by Eq. (1).

$$W_{ik} = (f_{ik}/AF) \times \log_2(n) - \log_2(AF) + 1 \tag{1}$$

where W_{ik} denotes weight of keyword k in i-th article, and f_{ik} denotes frequency of keyword k in i-th article. AF denotes frequency of article containing the specific keyword k in news video in processing, and n denotes the number of all articles in news video in processing. The reason why the number of all articles in the news video is used instead of the number of total articles in database, is because news articles are daily inserted to database. Eq. (1) is the normalized relative frequency in consideration of the number of all articles, the size of an article, and the frequency of keyword. Especially, in Eq. (1), the inverted document frequency in [6] and [7] is divided by the article frequency in order to assign large weight to keyword which has the high discriminating power. By this weighted indexing, the retrieval results about user's query can be ranked in descending order of the weights. After keyword is selected and weighted, inverted file which includes televising date, identification number of article, weight, news title, and name of broadcasting station of article having each keyword, is created.

4.3 Extraction of Color Feature in Representative Frame

For effective video browsing, we extract the frame which is able to represent an article after news video is segmented into articles. In this paper, we extract the frames synchronized with the first anchor, reporter, and interview tags as the representative frame. In order to perform the content-based retrieval to this representative frames, we extract feature vectors by using the color-structure descriptor of MPEG-7, and store these vectors in indexing database. Therefore, user is able to retrieve the frame in video by using not only keyword but also query by example. The color-structure descriptor (CSD) is a feature descriptor that captures both the color content and the structure of the content [8].

5 Video Retrieval System

The reuse of the large amount of video data is very important. The retrieval system searching faster and more accurately is required [9]. We design retrieval method, which consists of basic search and advance search. Basic search retrieves data by the position matched with string between the user specified query and the closed-caption. Advanced search retrieves data by extracting keyword from the closed-caption and performing weighted indexing. Also we are able to perform the content-based retrieval to representative frames of each segmented articles.

5.1 Basic Search

Basic search retrieves the voice-visual section synchronized with keyword that user queries. First, after query is input, the time synchronized with voice in video is achieved from database in Fig. 5. The start frame number for playing of voice-visual section is computed by Eq. (2).

$$F_{start} = (t - r) \times m \tag{2}$$

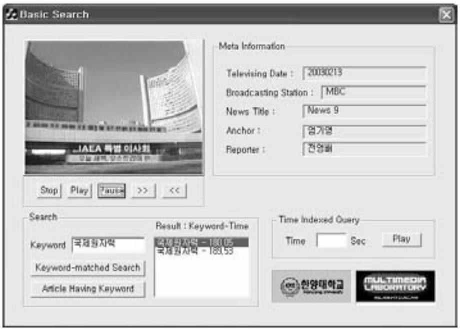
where F_{start} denotes the start frame number, t denotes the time synchronized with voice in video, and m denotes frame rates. r denotes the delay time of the closed-caption. r set to 3 in live broadcasting running with the real-time stenographic system, and r set to 0 in filmed broadcasting because caption is already synchronized with voice without the delay time.

In basic search, it is possible to search articles, which includes the query. First of all, searching the unit of article seeks the adjacent anchor tag that is received early than keyword matched with query. Then, the start position of article having the query is computed with the time synchronized with the anchor tag by Eq. (2). Also time indexed search, which plays from the time that user wants to search at, is included in the retrieval method. Fig. 8 illustrates a user interface of basic search.

5.2 Advanced Search

In the case of news video used as test-set in this paper, a large number of articles are created daily. In order to search the articles effectively, extraction for keywords having high discriminating power and weight-assignment of keywords are required, and also the use of metadata having the reporter's name, the televising date, and so on is required. Note that advanced search provides results ranked in descending order of keyword weight. User is able to perform basic search again within searched articles by the advanced search. Using the inverted file, searching by both favorite broadcasting station or news and the televising date is possible. Also, if user is data-manager in broadcasting station, information of anchor or reporter can effectively be used for searching article.

Fig. 9 shows an interface to search representative frames by query-by-example by using color-structure descriptor in MPEG-7. If user selects the representative


Basic Search

Meta Information

Televising Date : 20030213

Broadcasting Station : MBC

News Title : News 9

Anchor : 엄기영

Reporter : 전영배

Search

Keyword 국제원자력

Result : Keyword-Time

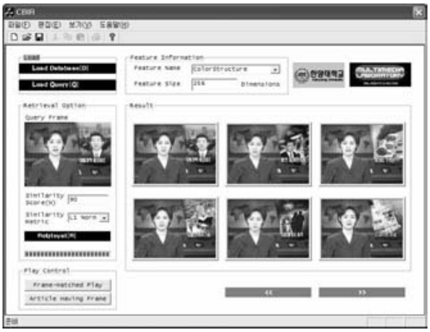
국제원자력 - 180.05

국제원자력 - 189.53

Time Indexed Query

Time Sec

Fig. 8. User interface of basic search


CBIR

Feature Information

Feature Name ColorStructure

Feature Size 256 Dimensions

Retrieval Option

Query Frame

Similarity Score(%) 90

Similarity Metric L1 Norm

Result

Fig. 9. Content-based retrieval using the color-structure descriptor

frame, user can retrieve metadata related to articles matched with the representative frame. Also user can play from the representative frame as a start frame or play all section of the article included in the representative frame.

6 Conclusion

In this paper, we have proposed the automatic indexing method of Korean closed-caption for knowledge-based video retrieval and the retrieval scheme using the indexed database. The time-alignment method has analyzed accuracy and delay time of the closed-caption generated by Korean stenographic system and synchronized voice data with the closed-caption successfully. The automatic indexing system has performed the morphological analysis and the disambiguation of the results. After processing of unknown-word, complex-noun, and stopword, the keywords have been selected as final index terms for retrieval.

Basic search of the proposed retrieval method has been easily able to search the voice-visual section synchronized with query and has played articles including key-word. Advanced search has been able to perform the search by using names of broadcasting, news, reporter, anchor, and televising date. The results of advanced search have been ranked in descending order of the weights of the matching key-words between query and articles. Also we have been able to perform the content-based retrieval to image frames, such as representative frames.

In experiment result, we have empirically confirmed that the closed-caption expressing voice naturally has been used for video retrieval. However, there is only 10∼20% broadcasting having the closed-caption for the hearing impaired-people recently in Korea. For not only welfare for the hearing impaired-people but also effectively storing and retrieving broadcasting data, the rate of closed-caption service needs to be increased regardless field of broadcasting.

In the future work, we are going to extend into documentary video, which includes a large amount of information of voice, and then apply to drama or movie. Also, we will apply speech recognition to the our system.

Acknowledgement

This work was supported by HY-SDR Research Center at Hanyang University, Seoul, Korea, under the ITRC Program of IITA, Korea.

References

1. Vittorio C., Lawrence D.B.: Image Databases. Jon Wiley & Sons, Inc. (2002) 261–279
2. Howard D.W., Takeo K., Michael A.S., Scott M.S.: Intelligent Access to Digital Video: Informedia Project. IEEE Computer (Digital Library Initiative special). Vol. 29 No. 5 (1996) 46–52
3. Alexander G.H. and Michael J.W.: Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval. Intelligent Multimedia Information Retrieval. AAAI Press/The MITPress. (1997) 215–240
4. Inderjeet M., David H., Mark T.M., Morgan G.: Towards Content-Based Browsing of Broadcast News Video. Intelligent Multimedia Information Retrieval. AAAI Press/The MITPress. (1997) 241–258
5. J. Cho: An Automatic Indexing System for Intelligent Information Retrieval. Master's thesis. Hanyang University. (1998) 26-28
6. William B.F., Ricardo B.-Y.: Information Retrieval: Data Structure and Algorithms. Prentice-Hall (1992)
7. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison Wesley Publishing. (1989)
8. MPEG-7 Visual Experimentation Model (XM), Ver. 10.0. ISO/IEC/JTC1/SC29 /WG11, Doc. N4062 (2001)
9. Wei Q., Lie G., Hao J., Xiang-Rong C., Hong-Jiang Z.: Integrating visual, audio and text analysis for news video. Proc. of Int. Conf. on Image Processing. Vol. 3 (2000) 520-523

Classification and Skimming of Articles for an Effective News Browsing

Jungwon Cho[1], Seungdo Jeong[2], and Byunguk Choi[3]

[1] Department of Computer Education, College of Education,
Cheju National University, 66 Jejudaehakno, Jeju-si, Jeju-do, 690-756 Korea
jwcho@cheju.ac.kr

[2] Multimedia Laboratory, Department of Electrical and Computer Engineering,
Hanyang University, 17 Haengdang-dong, Sungdong-gu, Seoul, 133-791 Korea
kain@mlab.hanyang.ac.kr

[3] Division of Information and Communications, Hanyang University,
17 Haengdang-dong, Sungdong-gu, Seoul, 133-791 Korea
buchoi@mlab.hanyang.ac.kr

Abstract. In order to browse the news video effectively, classification and skimming of news articles are positively essential. In this paper, we propose the classification and skimming of articles for an effective news browsing. The classification method uses tags to distinguish speakers in the closed-caption. The skimming method extracts the representative sentence from the part of article introduced by the anchor in the closed-caption and the representative frames consisting of anchor frame, open-caption frames, and frames synchronized with high-frequency terms. In the experiment, we have applied the proposed classification and skimming methods to news video with Korean closed-captions, and have empirically confirmed that the proposed methods could support effective browsing of news videos.

1 Introduction

As computing power and electronic storage capacity have grown, the main retrieval subjects users are reaching are from text and image to video. Video data is massive and various. For obtaining the precise retrieval results of queries, we are researching on indexing, storing, querying, and showing retrieval results [1]. The first thing we have to consider here is to analyze the user's demand and represent it in the language of indexing. However, most of the researches on multimedia information retrieval are focusing on using some features which are extracted automatically from images. In these researches, the forms of user queries are restricted. If the application scope in an information retrieval system (IRS) is a specific one like trademark retrieval, a user's query may be used more effectively. However, in the case of general video retrieval, knowledge-based retrieval [2] requires understanding the meaning of video data. Especially, the news video used in this paper can effectively be retrieved by using the knowledge-based query

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 704–712, 2005.

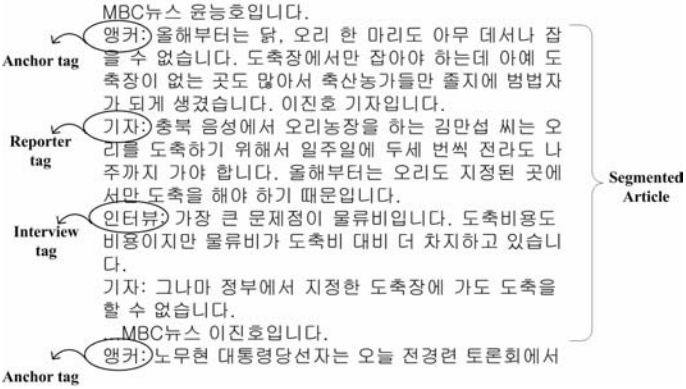
approach. This paper analyzes the meaning of news articles by using useful information extracted from the closed-captions. Also, we present classification and skimming methods for news articles using the closed-captions.

2 Classification

For classification, first, news video needs to be segmented into articles. Because the article is not a physical unit but a logical one, segmentation is difficult. This paper proposes a video segmentation method using closed-captions. The proposed method segments news video into articles by using tags in the closed-captions and extracts the reporter's name and the tags from the segmented articles. Then, the method successfully classifies those articles by using only the reporter's post database.

2.1 Segmentation of News Video Using Closed-Caption

The closed-caption has tags which are not contained in voice data. The tags are needed to discriminate who the speaker is. Generally, news video has three kinds of tags: ' **앵커:** ' (anchor), ' **기자:** ' (reporter), and ' **인터뷰:** ' (interview) tags as shown in Fig. 1. If we use the tags and structural characteristics of news video, we can achieve good performance in segmenting news videos with only the closed-captions.



MBC뉴스 윤능호입니다.
앵커: 올해부터는 닭, 오리 한 마리도 아무 데서나 잡
을 수 없습니다. 도축장에서만 잡아야 하는데 아예 도
축장이 없는 곳도 많아서 축산농가들만 졸지에 범법자
가 되게 생겼습니다. 이진호 기자입니다.
기자: 충북 음성에서 오리농장을 하는 김만섭 씨는 오
리를 도축하기 위해서 일주일에 두세 번씩 전라도 나
주까지 가야 합니다. 올해부터는 오리도 지정된 곳에
서만 도축을 해야 하기 때문입니다.
인터뷰: 가장 큰 문제점이 물류비입니다. 도축비용도
비용이지만 물류비가 도축비 대비 더 차지하고 있습니
다.
기자: 그나마 정부에서 지정한 도축장에 가도 도축을
할 수 없습니다.
...MBC뉴스 이진호입니다.
앵커: 노무현 대통령당선자는 오늘 전경련 토론회에서

Anchor tag
Reporter tag
Interview tag
Anchor tag
Segmented Article

Fig. 1. Tags in closed-caption: anchor, reporter, and interview

2.2 Classification of Articles Using Reporter Post Database

For an effective browsing of news articles, classification according to a category for the article is required [3]. In this paper, for classification, we extract the reporter's name in the closed-caption and match the name with records in the database having information of the reporter's posts. Consequently, the articles are classified into 7 categories.

We make the reporters post table - the lists of reporters in Munhwa Broadcasting Corporation (MBC) on Feb. 2003. Note that the news article is classified into one of 7 categories - politics, economy, society, unification, information science, culture, and international affairs - that are the same as classification categories in the MBC homepage. News articles consist of anchor, reporter, and interview parts. A position in the closed-caption, where the reporter's name appears, is fixed.

In Korea, news article is composed as follows; an anchor introduces the article briefly, and identifies the reporter's name. The reporter ends the article by referring to his or her name. By using that characteristic, we extract the reporter's name and query the reporter database. The search result of the query is the information of the reporter's post, which is able to indicate the category of the article. Fig. 2 is a flow-chart of the classification method using reporter's name that is extracted from the closed-caption of the article. In Fig. 2, two reporter's names are extracted from the two parts in an article. Two names are used for querying the reporter database. The results are names of the reporter's post. If the two names are the same, the article is classified into the category corresponding to the name of the reporter's post. However, if the results are different, we prefer the reporter's name that is extracted from the anchor part. Because there is no external noise in the broadcasting studio, the part from the anchor has less possibility of stenographic system error. Note that we prefer the part from the anchor.

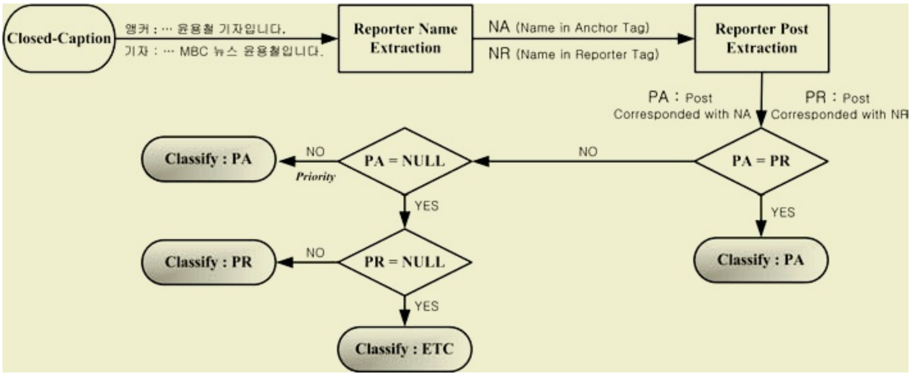


Fig. 2. Flow-chart of the proposed classification method

Before information retrieval of the news articles, we make the inverted file that is composed of information, such as broadcasting station, televising date, a start position for play, an end position, reporter name, and classified category.

3 Skimming

Playing all the news articles in order to confirm the validity of retrieval results is definitely a waste of time. In this paper, we propose the skimming method that summarizes the article with minimum loss of information. The result of the proposed skimming system consists of the representative sentences and the filmstrip having the anchor frame, the open-caption frames, and so on. At the same time, since the summarized result of the skimming system contains a filmstrip and text information, the distortion of the meaning can be minimized.

3.1 Extraction of Representative Frame

In this paper, the representative frames shown to users are the anchor frame, the open-caption frames containing the main contents of articles, and frames synchronized with high frequency terms. The anchor frame not only informs the starting location of an article but also has lots of the possibility to contain the window-caption. Therefore, an anchor frame is valuable as summary information. The previous research has already introduced the algorithm for extracting the anchor frame [4]. However, it needs many comparing operations. In this paper, we reduce operation times in extracting the anchor frame by using the anchor tag contained in the closed-caption. The anchor frame is continued over 10 seconds. If we extract the frame synchronized with the occurrence time of the tag, the extracted frame must be the anchor frame because the closed-caption and voice are completely synchronized. The time to extract the anchor frame is computed by the sum of the time at which the anchor tag appears and 3 seconds, as shown in Eq. (1).

$$F_A = (T_A + 3) \times R \tag{1}$$

Here F_A and T_A denote the frame number and the assigned time of the anchor tag respectively, and R denotes the frame rate. When the gradual scene change occurs, the previous article is continued with the anchor frame. Therefore we add 3 seconds in order to exclude the possibility of extracting the anchor frame in gradual scene change and extract the anchor frame including the window-caption.

This method of extracting frames synchronized with the closed-caption doesn't need the comparison-operation. If the synchronized frame is decoded, the frame is the accurate anchor frame. Fig. 3 shows the anchor frame extracted by the proposed method using the anchor tag in the closed-caption.

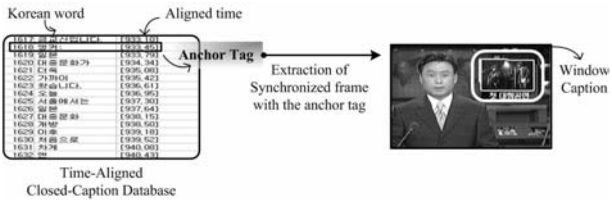

Korean word
Aligned time
Anchor Tag
Extraction of Synchronized frame with the anchor tag
Window Caption
Time-Aligned Closed-Caption Database

Fig. 3. Extraction scheme of the anchor frame using the anchor tag

In summarizing articles of news video, it is hard to convey the meaning of articles while only having general image data. However, if it extracts a frame containing the open-caption in the lowest part, the frame has sufficient information [5–8]. Note that the frame is chosen as a representative frame. As shown in Fig. 4, the open-caption has the sentence representing the article, the related place, the related person, or the name of the reporter. In this paper, for extracting the frame containing the open-caption effectively, we choose the I frames from the MPEG-1 video which is compressed data. With only I frames, it is enough to extract frames having open-captions because the open-caption is displayed for 1∼2 seconds. The feature of characters does not have hue information, so we use only luminance to extract features from MPEG data. We use the characteristics of the high frequency in the caption region. Besides, we exclude the DC coefficients to be robust regarding the illumination variation.

In order to extract the open-caption frame as the representative frame, we focus on the lowest part of frame as shown in Fig. 4.

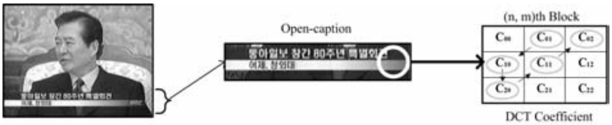


Fig. 4. Searching area for extraction and DCT coefficient

For the feature extraction, we use the block DCT coefficients in the search area and consider only five elements of the low frequency in the Zig-Zag scanning order. If we consider only lower frequencies ([1,0], [0,1]), some errors can be occur because general frames contain a lot of ingredients. That's why we select five elements. In this paper, we decide whether frames have the open-caption or not, using five elements ([1,0], [0,1], [1,1], [2,0], [0,2]) of low frequency in 8 × 8 block DCT coefficients. We describe the algorithm as the function $H_{n,m}$ which decides the feature of open-caption. Also we use a unit step function $u(c)$ to express the feature of character as a conditional formula.

For increasing the difference between the open-caption frame and common frames, we assign weight value α_i. To consider only coefficients, which is successfully able to reflect the characteristic of text, in DCT coefficients, we use two kinds of threshold values T_{high} and T_{low}. In this paper, we set T_{high} to 300, and T_{low} to 150 (experimentally determined).

$$\begin{aligned} H_{n,m} = \; & \alpha_1 \left[u(C_{10} - T_{high}) \times u(T_{low} - C_{01}) + u(C_{01} - T_{high}) \times u(T_{low} - C_{10}\right] + \\ & \alpha_2 \left[u(C_{11} - T_{high}) \times u(T_{low} - C_{20}) \times u(T_{low} - C_{02}\right] + \\ & \alpha_3 \left[u(C_{20} - T_{high}) \times u(T_{low} - C_{02}) + u(C_{02} - T_{high}) \times u(T_{low} - C_{20}\right] \end{aligned} \tag{2}$$

Fig. 5 shows the space meaning of each term in Eq. (2). (A) and (B) show the amount of low frequency's horizontal and vertical ingredient respectively, and (C) shows the amount of diagonal ingredient. Also, (D) and (E) express the amount of ingredients, such as the horizontal and vertical straight line of 8 × 8 block. n, m are block's index. That is, in the case that $H_{n,m}$ computed from the (n, m) block is over threshold, we consider that (n, m) block has the property of characters. It extracts only five elements of low frequency from DCT coefficients of each block in the lowest part of frame. Then, these coefficients are applied to Eq. (2).

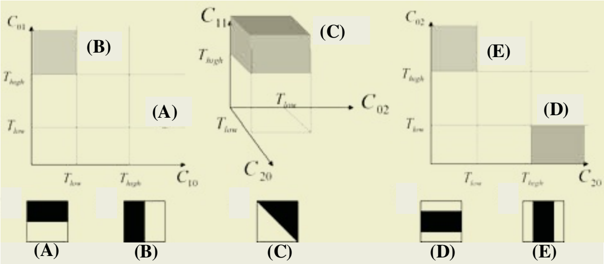


Fig. 5. Space meaning of each term in $H_{n,m}$

Each result from all blocks of a frame is accumulated. If the accumulated value is bigger than threshold ρ as shown in Eq. (3), we decide that the frame has the open-caption. Weights and threshold that we use are $\alpha_1 = 1, \alpha_2 = 3, \alpha_3 = 5$ and $\rho = 35$ (experimentally determined).

$$\sum_{n=0}^{N} \sum_{m=0}^{M} H_{n,m} \geq \rho \tag{3}$$

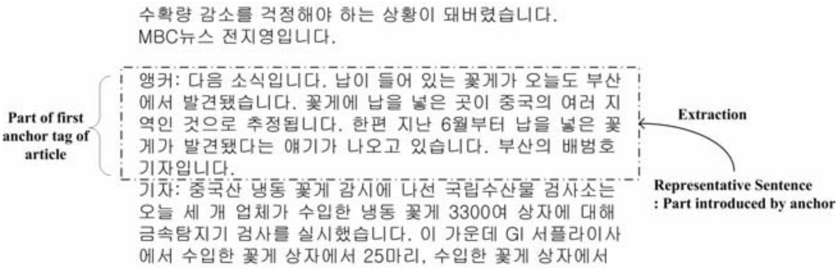
We are able to decide whether a frame is the open-caption frame or not without scanning all frames. In order to prevent extracting frames which have the same caption, the accumulated value $H_{n,m}$ in the frame decided as the open-caption frame is stored, and the accumulated value is compared with one computed from the next frame. As shown in Eq. (4), if the difference between the i-th frame and j-th frame is not over the limitation ϕ, we decide that two frames have the same open-caption. In this paper, we set ϕ to 5 (experimentally determined).

$$\left| \sum_{n=0}^{N} \sum_{m=0}^{M} H_{n,m}^{i} - \sum_{n=0}^{N} \sum_{m=0}^{M} H_{n,m}^{j} \right| < \phi \tag{4}$$

The proposed skimming system extracts 6 representative frames of an article. If the anchor frame and the open-caption frames are 5 frames below, the rest of the 6 frames are extracted by using term frequency. Extraction based on the term frequency, first, selects keyword from articles and computes frequency of each keyword. Then, higher ranked terms to fill the filmstrip are selected. When frames synchronized with the selected terms are extracted except for frames in the anchor's part, the first frame synchronized with the term in the reporter's part is extracted because the frame synchronized with the term in anchor's part is the anchor frame.

3.2 Extraction of Representative Sentence

In this paper, for effective browsing, the sentence which represents the article and the filmstrip which consists of representative frames are provided. The representative sentence, which is the best suitable, is the part where the anchor introduces the article. As shown in Fig. 6, we confirm that the anchor part conveys the abbreviated contents of the article.



수확량 감소를 걱정해야 하는 상황이 돼버렸습니다.
MBC뉴스 전지영입니다.

앵커: 다음 소식입니다. 납이 들어 있는 꽃게가 오늘도 부산에서 발견됐습니다. 꽃게에 납을 넣은 곳이 중국의 여러 지역인 것으로 추정됩니다. 한편 지난 6월부터 납을 넣은 꽃게가 발견됐다는 얘기가 나오고 있습니다. 부산의 배범호 기자입니다.

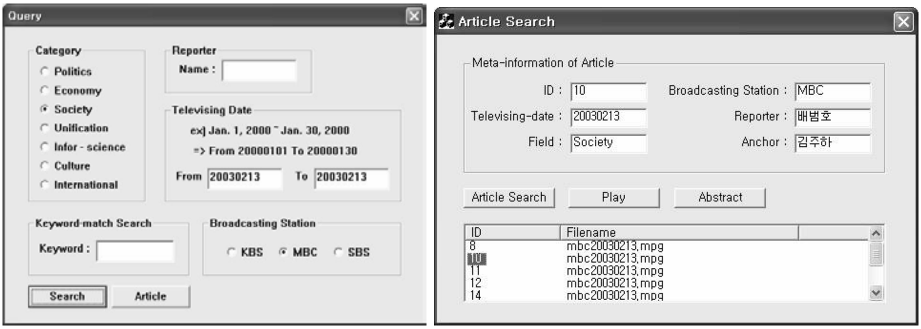
기자: 중국산 냉동 꽃게 감시에 나선 국립수산물 검사소는 오늘 세 개 업체가 수입한 냉동 꽃게 3300여 상자에 대해 금속탐지기 검사를 실시했습니다. 이 가운데 GI 서플라이사에서 수입한 꽃게 상자에서 25마리, 수입한 꽃게 상자에서

Part of first anchor tag of article
Extraction
Representative Sentence : Part introduced by anchor

Fig. 6. Extraction of representative sentence

4 Experimental Results

For the experiment, we use the news video broadcasted by MBC of Korea on Feb. 2003. The news video includes the closed-caption generated by Korean stenographic system. Fig. 7 shows the user interface for making a query and the retrieval result from the classified database. Before confirming content details of the article, user is able to check basic bibliographical information with metadata.



Query

Category
○ Politics
○ Economy
● Society
○ Unification
○ Infor - science
○ Culture
○ International

Reporter
Name :

Televising Date
ex] Jan. 1, 2000 ~ Jan. 30, 2000
=> From 20000101 To 20000130
From 20030213 To 20030213

Keyword-match Search
Keyword :

Broadcasting Station
○ KBS ● MBC ○ SBS

(a)

Article Search

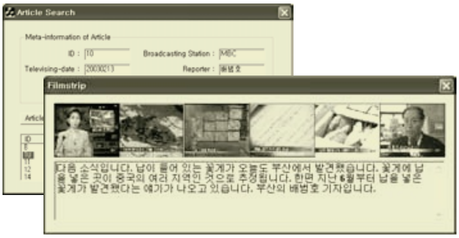
Meta-information of Article
ID : 10
Broadcasting Station : MBC
Televising-date : 20030213
Reporter : 배범호
Field : Society
Anchor : 김주하

ID	Filename
8	mbc20030213,mpg
10	mbc20030213,mpg
11	mbc20030213,mpg
12	mbc20030213,mpg
14	mbc20030213,mpg

(b)

Fig. 7. User interface for query (a) and search result (b)

Fig. 8 shows a skimming result consisting of representative frames and representative sentence. We are successfully able to browse news articles by using the skimming result as shown in Fig. 8.



ID : 10
Broadcasting Station : MBC
Televising-date : 20030213
Reporter : 배범호

다음 소식입니다. 납이 들어 있는 꽃게가 오늘도 부산에서 발견됐습니다. 꽃게에 납을 넣은 곳이 중국의 여러 지역인 것으로 추정됩니다. 한편 지난 6월부터 납을 넣은 꽃게가 발견됐다는 얘기가 나오고 있습니다. 부산의 배범호 기자입니다.

Fig. 8. Example of skimming result

In order to justly evaluate the performance of the classification of news articles, we compare the classified results provided by MBC news VOD server (http://imnews.imbc.com) with the classified results of our proposed system about 259 articles, and calculate the precision ratio as shown in Table 1. The reason of the principal error shown in Table 1 is short articles which don't have the reporter's name. If short articles are excluded, the precision is 95%.

Table 1. Precision ratio of the classification method

The number of total articles	T	259
The number of articles having post information	D	246
The number of correctly classified articles	C	235
Precision ratio for all articles	C/T	90.7%
Precision ratio for articles having post information	C/D	95.0%

In order to evaluate the proposed skimming system, we want to verify the methods for selecting representative sentence and frames. The extraction method of the anchor frame, frames synchronized with the frequently appeared terms, and the representative sentence is straightforward. Therefore, we confirmed the performance of the module which extracts the open-caption frame using the proposed algorithm. Table 2 shows the precision of extraction of open-caption frames with 66 open-caption frames.

Table 2. Precision ratio of extraction of open-caption frames

The number of total open-caption frames		66
The number of the extracted frames	F	62
The number of correctly extracted frames	H	52
The number of false-alarm frames		5
The number of false-dismissal frames		4
The number of duplicative extracted frames		1
Precision ratio	H/F	83.9%

Even though there were two causes of error that the lowest part has some characters which are not the open-caption, and the extracted frame has the same open-caption as the previous frame, the result shows precision of 83.9%. Note that, as a result, it is suitable for the skimming system.

5 Conclusion

In this paper, for effective browsing technique, we have proposed the classification and skimming method of articles for an effective news browsing. The experimental result of the classification method is almost the same with the result of the handmade classification. The skimming method is able to provide the filmstrip and representative sentences which are able to reflect meaning of a news article, so we have confirmed that the proposed method is suitable for knowledge-based retrieval which can reflect the user's demand. The proposed methods overcame the drawback that previous techniques could hardly convey the meaning in content-based retrieval using the features of color, shape, texture, and so on. Moreover, since the processing mechanism in this paper is easy and runs quickly, it is believed that the result can be employed in on-line applications. We are now working on extracting the representative frames by using information, such as the camera motion, and so on. Also we are now working on abstracting representative sentences by the natural language processing of the closed-caption.

Acknowledgement

This work was supported by HY-SDR Research Center at Hanyang University, Seoul, Korea, under the ITRC Program of IITA, Korea.

References

1. Vittorio C., Lawrence D.B.: Image Databases. John Wiley & Sons, Inc. 2002
2. Alexander G.H., Michael J.W.: Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval. Intelligent Multimedia Information Retrieval. AAAI Press/The MIT Press. (1997) 215–240
3. Ariki Y., Matsuura K.: Automatic classification of TV news articles based on telop character recognition. IEEE International Conference on Multimedia Computing and Systems. Vol. 2. (1999) 148–152
4. HongJiang Z., Yihong G., Smoliar S.W., Shuang Y.T.: Automatic parsing of news video. in Proc. of the Int. Conf. on Multimedia Computing and Systems. (1994) 45–54
5. Smith M.A., Kanade T.: Video skimming and characterization through the combination of image and language understanding. International Workshop on Content-Based Access of Image and Video Database. (1998) 61–70
6. HongJiang Z., Chien Y.L., Stephen W.S., JianHua W.: Video Parsing, Retrieval and Brows-ing: an Integrated and Content-based Solution. Intelligent Multimedia Information Retrieval. AAAI Press/The MIT Press. (1997) 139–158
7. Girgensohn A., Boreczky J.: Time-constrained keyframe selection technique. IEEE Interna-tional Conference on Multimedia Computing and Systems. Vol. 1. (1999) 756–761
8. Christel M.G., Hauptmann A.G., Warmack, A.S., Crosby, S.A.: Adjustable filmstrips and skims as abstractions for a digital video library. ADL '99. Proceedings. IEEE Forum on Research and Technology Advances in Digital Libraries. (1999) 98–104

Intelligent Tutoring System with 300-Certification Program Based on WIPI

Youngseok Lee[1], Jungwon Cho[2], and Byunguk Choi[3]

[1] Department of Electrical and Computer Engineering, Hanyang University,
17 Haengdang-dong, Sungdong-gu, Seoul, 133-791 Korea
yslee38@mlab.hanyang.ac.kr

[2] Department of Computer Education, Cheju National University,
66 Jejudaehakno, Jeju-si, Jeju-do, 690-756 Korea
jwcho@cheju.ac.kr

[3] Division of Information and Communications, Hanyang University,
17 Haengdang-dong, Sungdong-gu, Seoul, 133-791 Korea
buchoi@mlab.hanyang.ac.kr

Abstract. Recent developments in content processing technology and the widespread diffusion of wired and wireless Internet mean that users can now learn by means of a computer, anytime and anywhere. English learning under the multimedia environment is able to increase the interest of learners and lead to the development of their communication ability. Although using computers to teach English in a conventional educational environment provides motivation and effective learning on the part of the students, the method still has problems, which include the provision of learning materials without consideration of teaching methods, and evaluation without provision for differences in individual student levels. In order to solve the problems and take the advantages, we propose the Intelligent Tutoring System (ITS) for English learning with mobile technology. With overcoming limitations of the mobile environment and using proper mobile contents, the proposed system provides the effective learning method based on the ITS which is able to replace teacher's roles.

1 Introduction

Recent technologies of content delivery systems through the Internet enable users to access the learning content in anywhere and anytime. The conventional instructional methods in a face-to-face learning environment have helped and facilitated learning for learners. However, the methods have been criticized for not being able to provide just-in-time supports when the learner's achievement is fluctuated that he or she demonstrates successful learning outcomes but sometimes not [1]. This type of various learning pattern may be happened when an individual learner does not have deep understanding to a given learning task. Without having deep understanding, the learner's learning outcome can't be reliable to a variety of situation. The unreliable understanding could be more a serious problem for an effective instruction if there is a group of learners to be taught [2, 3]. In addition there has, so far, been little research focused on learning or education methods involving mobile devices [4-6].

Various factors, including the importance of particular subjects and the degree of difficulty must be considered when seeking to provide content suitable for specific

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 713–720, 2005.

student levels; it is necessary to estimate each student's level and then provide appropriate learning materials. Learning materials must be ranked according to the degree of difficulty, as well as with regard to levels of distinction and weighting values, something that is not necessary for other methods of educational evaluation, such as Item Response Theory (IRT). The need to accommodate a variety of potential learning devices (cellular phones, PDA) also complicates the design of a useful system.

Our proposed system is based on an ITS that provides suitable content for specific student levels in a multi-modal education platform. We describe the subject contents and create a learner model that determines the provision of material appropriate to specific student levels, as assessed by an inference engine based on IRT and learner's model. After learner completes those test items, the system will give feedback from result instead of teacher's role. A chain of processing leads to interaction between teacher and student, increases the efficiency of learning, and supports teacher to instruct and give a feedback.

2 Literature Review

2.1 Item Response Theory (IRT)

In order to estimate a student's ability it is necessary, initially, to conduct tests; the use of item response theory enables an assessment of parameters such as item difficulty and item surmise. Item difficulty means the degree of difficulty of an item; the IRT formula for calculating item difficulty is as follows (1) [7].

$$b = \frac{R}{N} \tag{1}$$

B: Item difficulty, *N*: Number of total examinee, *R*: Number of a correct answer

Cangelosi defined a difficult item as one for which B<0.25, a suitable item as one for which 0.25<B<0.75, and an easy item as one for which B>0.75 [7]. The degree of surmise assigns a numerical value to the number of students answering correctly by guessing. The IRT formula for calculating item surmise is as follows (2).

$$C = \frac{W}{N(Q-1)} \tag{2}$$

C: Item difficulty, W: Number of examinees answering incorrectly,
N: Total number of examinees, *Q*: Number of possible answers

Item surmise values range from 0 to 1. In the case of very difficult items, item surmise values higher than item difficulty values can be observed [7].

2.2 Intelligent Tutoring System (ITS)

Changing teaching methods in Computer Assisted Instruction (CAI) systems is a difficult task. ITS, however, is based on artificial intelligence (AI) and supports an adaptable and dynamic learning environment, which overcomes the limitations of CAI [8]. Generally, ITS consist of an interface module, a learner module, an instruction module and an expert module. The expert module builds and applies the obtained knowledge base through pertinent professional advice. It also manages information about suitable learning objectives, material, and specific instruction conditions [9].

The learner model shows the current state of a student's knowledge and the diagnostic processor infers the learner model. A student's level is computed by diagnostic function, which divides the current state of the student's knowledge [8]. An Instruction module decides what to teach, when to teach, and how to teach, given the learning status of the student [9]. The interface module provides an interface through which student and system can interact. It is important that the interface module offers a friendly and appropriate interface and can respond to interactions with the student.

2.3 Wireless Internet Platform for Interoperability (WIPI)

WIPI is the standard specifications for the mobile platform made by the Mobile Platform Special Subcommittee of the Korea Wireless Internet Standardization Forum (KWISF). These are standard specifications necessary for providing an environment for mounting and implementing applications downloaded via the wireless Internet on the mobile communication terminal. As a Mobile Platform Special Subcommittee of KWISF, the WIPI Forum receives the requirements for writing WIPI specifications from mobile communication companies, mobile platform developers, terminal manufacturers, and content developers and performs standardization activities [10].

3 Design of System

3.1 Reconstruction of 300-Certification Program for English Conversation

The 300-Certification program for English conversation is currently used in Korean primary schools. The teaching material consists of 300 sentences organized into groups. To distinguish learning ability, the evaluation of listening function is based on multimedia content and includes character discernment of sound, stress and intonation, conservation ability of learned contents, etc. According to a systematic design of instruction [11, 12], the 300-Certification program contents resulted in a classification scheme involving four types of item useful for general evaluation Table 1.

Table 1. Types of item and form

Types	Example
A	Listen, and Choose a suitable Picture
B	Listen, and Choose a Correct Answer
C	Listen, and Choosing a missing statement
D	Listen, and Choosing a including statement

3.2 System Architecture

This paper proposes a system that consists of an expression section that presents learning materials to the student, the ITS core for data processing and management, and sections involved in forms of communication. Fig. 1 shows the architecture.

The communication module provides learning contents and materials to web and cellular phones for CDMA. The ITS core consists of an inference engine that chooses materials appropriate for the student and studies the student's learning progress. The ITS core was designed logical formulation of learning to measure the learner's trait and the learner's knowledge, and acts on basis of inference rule [13].

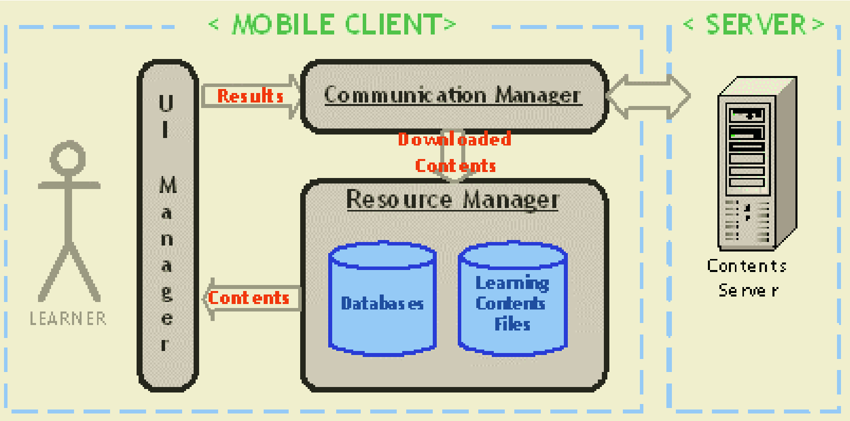


Fig. 1. Architecture of system

Interface module consists of the input interpretation and the output creation system. Input interpretation recognizes input of learner through keypad in the cellular phone and then analyzes the agreement between a correct answer and input data. Output creation present sound contents through speaker of the cellular phone, and image contents through the LCD screen.

Teacher module set up the strategy of instruction by estimating value of learner's level and database of learner's model. The strategy of instruction decides to provide suitable items of learner's level and prefer to instruction method, classification of items disposition.

Expert module provides knowledge into other module. Knowledge based of expert module is a hearing domain in English of elementary school, source of knowledge is a textbook and teacher's manual which of each grade in English of elementary school. Fig. 2 shows the core mechanism for the proposed ITS.

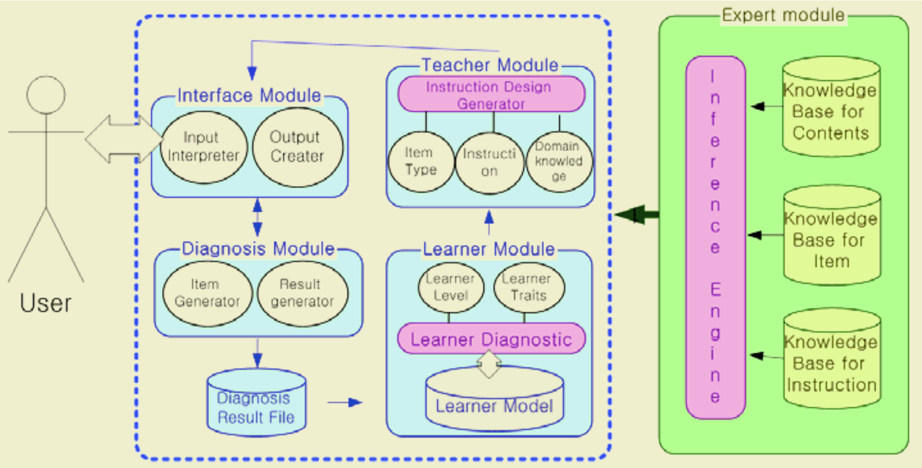


Fig. 2. Core mechanism of the proposed ITS

3.3 Creation of Student’s Learning Model

The proposed system creates learning models to display the student’s present knowledge state. According to the knowledge engineering process [13], we define item Type Diagnostic Value (TDV) that explicates the learning model to diagnoses a student's learning process for each item type in the following way.

$$TDV = \frac{\sum_{i=1}^{n} QW_i}{\sum_{i=1}^{n} QA_i} \tag{3}$$

i: Total Number oh Sheets, QA: Total Number of Items,

QW: Number of examinees answering incorrectly

TDV is set a problem as a numerical formula based on ratios. Ratios are calculated by comparing the total number of item type index to the number of items with an incorrect response. This kind of analysis can supply feedback on learning progress, uncover the weak points of each student’s understanding.

4 Implementation of System

This system was implemented in the environment described in Table 2.

Table 2. Implementation environment

	Contents Server	Mobile Client
H/W	Intel Pentium 3 1.0 GHz Dual CPU 512 MB RAM	SKT IM-6100 256 Color Display, 16 poly Sound
S/W	Windows 2000 Server MS SQL-2000, ASP	Intel Pentium 3 1.0 GHz Dual CPU 512 MB RAM

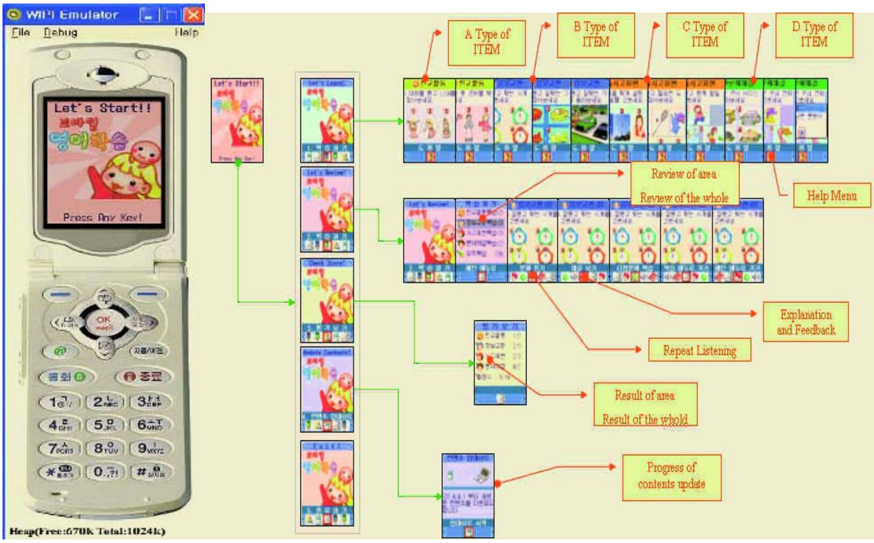


Fig. 3. Result of implementation

Fig. 3. shows the result of implementation. The system provides tests after the completion of study units. This system functions through response to menus. Menus present evaluation items that are presented in groups of four types that reconstruct curriculum contents. If any student selects the review menu, he reviews previous learning contents or could see the result about each item and feedback. Update menus could get new contents through images, audio content and/or dialog for each curriculum section and English sentence. If you want to demonstrate, connect our site - http://andy.hanyang.ac.kr/english.html - and see mobile version and web version.

5 Evaluation of System

This system was applied to Elementary Schools in Seoul, with about 100 students in 6 months. The results are shown in Table 3.

Table 3. Result of analysis for 208 items

Type A			**Type B**		
Item ID	Difficulty(Surmise)		Item ID	Difficulty(Surmise)	
	Inference	Normal		Inference	Normal
A601-0001	0.71(0.10)	0.43(0.19)	B601-0001	0.80(0.07)	0.57(0.14)
...	...	...	...	...	...
A626-0002	0.79(0.07)	0.45(0.18)	B626-0002	0.78(0.07)	0.25(0.25)
Type C			**Type D**		
Item ID	Difficulty(Surmise)		Item ID	Difficulty(Surmise)	
	Inference	Normal		Inference	Normal
C601-0001	0.75(0.08)	0.30(0.23)	D601-0001	0.74(0.09)	0.55(0.15)
...	...	...	...	...	...
C626-0002	0.48(0.17)	0.50(0.17)	D626-0002	0.85(0.05)	0.50(0.17)

First of all, we analyzed performance of inference engine by Maximum-likelihood learning in statistical learning method. In case of made use of inference engine in all types, appeared about 0.3 points difficulty higher. It shows that in case of inference yes, 3 people among 4 people select correct answer, other case, 2 people among 4 people select correct answer. It can assume that become learning by reasoning successfully that 1 person can select more correct answer among 4 learners.

Students could be divided into groups based on whether or not they used the inference engine Table 4. You have regard that one sheet include 10 items.

Table 4. Testing result by using inference engine

	Repeat Testing		**Learning by Feedback and Repeat Testing**	
	Inference	Normal	Inference	Normal
Average Solved Sheets	10	11	12	34
Average Score	3.9	6.0	7.5	7.3

In repeat testing, Groups not using the inference engine scored, on average, 2 points less than those using the inference engine. Performance of inference engine dropped in repeat testing. In case of repeat learning by feedback and testing, performance was showed. However, through feedback to the inference engine through longer periods of learning and testing, it could be expected that its ability to improve the learning levels of students would increase.

6 Conclusion

Existing web-based education systems focus on providing study materials without considering the knowledge status or ability of individual students. This paper proposes an intelligent tutoring system that provides materials appropriate for specific student levels via a mobile education platform that supports various communication environments and devices.

This paper have designed and implemented the system to conduct the appropriate education according to the learner's level in mobile environment and to obtain feedback of the learning. Installation of such a system in elementary schools, as part of the English teaching program, showed that the inference engine can provide materials appropriate to the level of each individual student.

Acknowledgment

This work was supported by HY-SDR Research Center at Hanyang University, Seoul, Korea, under the ITRC Program of IITA, Korea.

References

1. Davidovic A., Warren J., and Trichina E.: Learning Benefits of Structural Example-Based Adaptive Tutoring Systems, IEEE Trans. on Education 46 (2003) 241-251
2. Vasileva T., Trajkovic V., Cabukovski V., and Davcev D.: An Algorithm for Expert Level Estimation in a Distance Educational System, 2nd Asia-Pacific Forum on Engineering & Technology Education (1999) 346-350
3. Vasileva T., Trajkovic V., and Davcev D.: Experimental data about knowledge evaluation in a distance learning system, IFSA World Congress and 20th NAFIPS International Conference IEEE (2001) 25-28
4. Phanwoo Park: Distance Education System for English Learning on Internet, Frontiers in Eduation Conference, FIE 2 (1998) 760-765
5. Koyama A., Sasaki A., Barolli L., and Cheng Z.: An agent based education system for cellular phone, IEEE 12th Int. Workshop on Database and Expert Systems Applications (2001) 198-202
6. Chi-Hong LEUNG, et al.: Mobile Learning: A New Paradigm in Electronic Learning, Proceeding of the 3rd IEEE International conference on Advanced Learning Technologies (2003) 76-80
7. Crocker L. and Algina J.: Introduction to classical and modern test theory, Holt, Rinehart and Winston, inc. (1996)

8. Badjonski M., Ivanovic M., and Budimac Z.: Intelligent tutoring system as multiagent system, Proceeding of IEEE Trans. on ICIPS 1 (1997) 28-31
9. Okamoto Toshio,: The Current Situations and Future Directions of Intelligent, IEICE Trans. on Information & System E77-D (1994) 143-161
10. WIPI Forum, http:// wipi.or.kr/English/index.html (2004)
11. Walter D., Lou C., and James O. C.: The Systematic Design of Instruction, 5th edn. Pearson Education, Inc. (2001)
12. KERIS, http://www.edunet4u.net (2004)
13. Russell S., Norvig P.: Artificial Intelligence - A Modern Approach, 2 edn. Pearson Education, Inc. (2003)

ECA Rule Based Timely Collaboration Among Businesses in B2B e-Commerce

Dongwoo Lee[1], Seong Hoon Lee[2], and YongWon Kim[3]

[1] Department of Computer Science, Woosong University
17-2 Jayang-dong Dong-ku, Daejon 300-718 Korea
dwlee@woosong.ac.kr

[2] Department of Computer Science, Chonan University
115 Anseo-Dong, Chonan, Choongnam 330-180 Korea
shlee@mail.chonan.ac.kr

[3] Department of Information Management, Konyang University
ywkim@konyang.ac.kr

Abstract. In this paper collaboration among businesses in B2B E-commerce is analyzed and the need for timely collaboration is derived and classified in terms of inter-organizational contracts. To meet the need a method of event-condition-action (ECA) rule based timely collaboration and an intelligent active functionality component (IAFC) are proposed to provide B2B E-commerce systems with flexible coordination and timely processing in WWW environment. The proposed method supports high level programming and event-based processing so that system administrators and programmers can easily maintain the timely collaboration independently to the application logic.

1 Introduction

B2B E-commerce systems need to be coordinated and integrated for collaboration inter-organizations to achieve their common business goals. Especially emergency requests or critical information among businesses should be processed in an immediate mode. Most current systems, however, due to the systems' security and autonomy, cannot handle these requirements appropriately, but handle them in a batch processing mode or ad hoc manners [1].

In this paper collaboration among businesses in B2B E-commerce is analyzed and the need for timely collaboration is derived and classified in terms of inter-organizational contracts. To meet the need a method of event-condition-action (ECA) rule [2] based timely collaboration and an intelligent active functionality component (IAFC) are proposed to provide B2B E-commerce systems with flexible coordination and timely processing in WWW environment. Since high-level ECA rule programming is supported, the collaboration among businesses and event-based immediate processing can be implemented independently to application logic. Thus, system administrators and programmers can easily program and maintain timely collaboration among businesses in B2B E-commerce.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 describes the need for timely collaboration among businesses in B2B EC, the proposed mechanism to meet it, and requirements to provide the mechanism. In

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 721–727, 2005.

section 4, one of the requirements, elements of an ECA rule program and examples of ECA rule programs are presented. Section 5 discusses architecture of the IAFC and implementation and evaluation of a pilot system is described. Finally in section 6, we conclude with future work.

2 Related Work

The advent of the internet, WWW, and distributed computing technologies has been enabled business organizations to conduct business electronically. And a lot of researches on B2B E-Commerce have been carried on [3]. But the most of researches have been mainly focused on interoperability problems among businesses. The issues of timely collaboration among businesses have not been addressed comprehensively.

There are many researches on exception handling issues on business processes [4, 5, 6]. The exception is defined as deviation from the normal workflow, such as system errors or failures that interrupt normal processing of workflows. The exceptions are classified into basic failure on system level, application failure, expected exception on workflow level, and unexpected exception. Especially [4, 6] propose ECA rule based exception handling methods. However, these researches are focused on normal processing of workflows with the exceptions, i.e., fault-tolerant workflow processing. That is different from timely collaboration issues in this paper.

In summary, previous work addressed either general interoperability issues of businesses or exception handling of business processes' failures. Few of them comprehensively address timely collaboration issues among businesses in B2B E-Commerce.

3 Timely Collaboration Among Businesses

3.1 The Need for Timely Collaboration

Consider a typical B2B E-Commerce shown in Fig.1 as a motivating example. All participants are connected by solid lines which denote internet and information flow. The dotted arrows denote material flow among participants. Each participant does business with each other by its own B2B and B2C systems. That is, a shopping mall takes customers' orders and provides services by B2C system, while it places orders to suppliers, requests delivery of items to shippers and money transfer to banks, inquires customers' credit to credit card company based on the agreement or contracts which were made with other participants. A supplier receives orders from shopping malls by its B2B system.

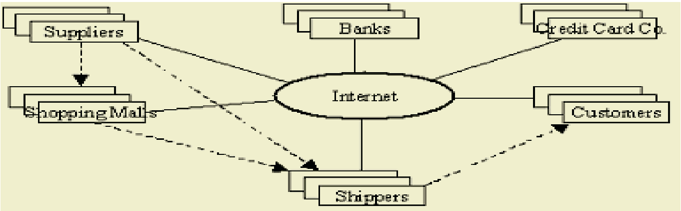


Fig. 1. Typical B2B E-Commerce

In the above B2B EC there are two kinds of job processing modes in businesses. One is a batch-processing mode in which a business collects jobs and processes them at a time. The other is an immediate processing mode in which a business processes jobs promptly when they come in. In the former mode human or systems work efficiently while the processing time of each job becomes longer. In the latter mode human or systems should always wait for a job while each job is processed promptly. The choice of processing modes depends on the characteristics of jobs, policy of a business, and contracts between businesses.

Timely collaboration among businesses, i.e. immediate request - immediate cooperation, can be seen as exceptions out of business' normal collaborations. That is, emergency request or critical information among businesses should be transmitted to partners promptly and processed by the partners in an immediate mode. They are not frequent, but once they occur they may require special treatment and affect customers' or businesses' profits in a large degree. Since businesses collaborate each other by contract fulfillment, the cases for the timely collaboration can be classified as following in terms of service contracts:

1. unable to fulfill a normal contract service
2. need to modify or compensate a normal contract service
3. need to cancel a normal contract service
4. need a special service instead of a normal contract service

For the above cases, new contracts, which require timely collaboration, can be added into systems incrementally. Most current systems, however, due to the systems' security and autonomy, cannot handle these requirements appropriately, but handle them in a batch processing mode or *ad hoc* manners [1]. That is, they use login method by allowed users or Email. Or low-level *ad hoc* programs, which are coded into application logic, handle the cases. It causes software modularity problems.

3.2 Timely Collaboration by ECA Rule Paradigm

We derived the timely collaboration procedure among businesses in B2B EC shown in Fig. 2, which consists of 4 phases:

1. Detection: a phase to detect that a business wants to make emergency requests for partner's cooperation or critical information occurs, which should be transmitted promptly.
2. Transmission: a phase to transmit the detected situation to a partner promptly.
3. Evaluation: a phase to evaluate collaboration constraints whether they should be processed in an immediate mode. There are two kinds of constraints, time constraints and resource constraints.
4. Processing: a phase to execute or process the requested job or the information in an immediate mode.

As shown above, in order to collaborate in an immediate mode, the situation for timely collaboration should be detected, notified or transmitted to each other, evaluated and recognized, and processed promptly. It shows that timely collaboration among businesses is suitable application to ECA rule mechanism [2]. That is, the timely collaboration can be represented in ECA rule, such as occurrence of the situation for timely collaboration as event, collaboration constraints as condition, and the processing of the job as action. Then, an IAFC, which process ECA rules, detects automatically the occurrences of the event and notify the occurrence to partner's system. The IAFC of the partner evaluates the condition. If the condition is satisfied then it executes the action promptly for collaboration. That is, the collaboration among businesses can be processed in an immediate mode without interference of application or users.

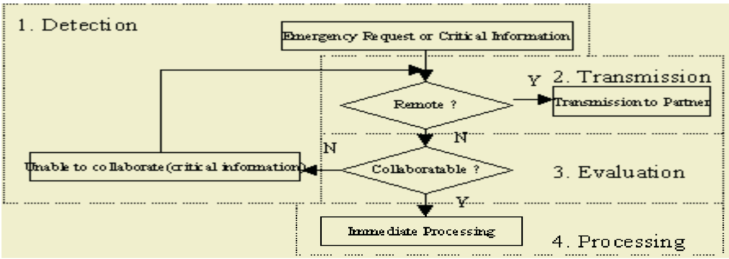


Fig. 2. Timely Collaboration Procedure among Businesses in B2B E-Commerce

The timely collaboration among businesses in B2B EC can be supported by the two main elements, ECA rule programming facility and intelligent active functionality component (IAFC). In the next sections ECA rule programming and IAFC are described.

4 ECA Rule Programming

In this research new ECA rule language is not developed instead existent ECA rule language is extended at the minimum to express timely collaboration. We adopt ECAA (Event Condition Action Alternative Action) rule pattern, which is one of ECA rule patterns [2]. Fig. 3 depicts ECA rule structure. Alternative Action part is for flexible representation of timely collaboration. Even though a business makes a timely collaboration request, its partner cannot cooperate because of the collaboration constraints. In this case the partner should notify the inability for cooperation to requester as an alternative action. Therefore it is optional.

The rule-name does not have a major role in its execution since it is triggered by an event. Rule names are used mainly for management purpose.

Rule rule-name;
  Event event-expression;
  Condition condition-expression;
  Action Begin action-block **End**
  [**Alternative Action Begin** alternative-action-block **End**]

Fig. 3. ECA Rule Structure

Event: A rule is triggered by detection of occurrence of an event described in event-expression. In this paper the events are classified into local and remote in terms of occurrence and subscription. If local event occurs and is subscribed by a remote system, it is transmitted to the system. This information is registered into Event-Schema-

Table of Event-Manager in IAFC when it is defined. In terms of contents, the events are furthermore classified as request event and notification event as in [Table 1].

Table 1. Events and Actions for Timely Collaboration

Request Events & Notification Events	Timely Collaborative Action
Notify-Able-Service	No-Action
Notify-Unable-Service	Find-Alternate-Service
Request-Modify-Service	Modify-Service
Request-Cancel-Service	Cancel-Service
Request-Special-Service	Special-Service

Condition: Once an event has been detected, a condition part is evaluated. For timely collaboration, collaboration constraints should be checked. Even though a business requests timely collaboration, its partner may not be able to cooperate. There are two kinds of constraints. One is time constraints that the partner should provide requested service within a contracted time period. The other is resource constraints that the partner should have resources such as man power, required system, or items to provide requested service.

Action: If the condition is satisfied, the action block is invoked. An action block consists of a set of actions or call statements, which execute requested service for timely collaboration. [Table 1] shows kinds of events and corresponding action types.

Example of ECA Rule Program: Consider that a shopping mall becomes short of an item suddenly and requests a partner supplier to provide it. Then the supplier should provide the item quickly within a contracted time period. If, however, the item is out of stock in the supplier's warehouse, it should be notified to the shopping mall promptly so that the shopping mall can try to find alternate supplier to fill the item.

This example shows that a shopping mall and a supplier should collaborate in an immediate mode. It can be implemented by the following two rules:

```
Rule Find-Alternate-Service /* rule on a Shopping Mall */
        Event unable-special-supply (string supplier, string item-1, integer n);
        Condition true;
        Action Begin find-alternate-service(string item-1, integer n)
             End

Rule Special-Service /* rule on a Supplier */
        Event request-special-supply(string requester, string item-1, integer n);
        Condition no. of item-1 > n;
        Action Begin special-order-processing(string requester, string item-2) End
        Alternative Action Begin raise-event('notify-unable-special-supply') End
```

5 Intelligent Active Functionality Component

Timely collaboration written in ECA rules is processed by an IAFC. The architecture of an IAFC consists of five modules, Communication-Manager, Event-Manager, Rule-Manager, Event-Rule-Interface, and Action/Application. The overall architecture is shown in Fig. 4. Detailed explanation of design and implementation of the IAFC refers to [7]. Brief description is following:

Communication-Manager: It is implemented in Java servlet of Web server and contains two roles. First it receives XML message through Web server, extracts event, and transfer to local Event-Manager. Secondly it receives event from local Event-Manager, transforms into XML format, and using HTTP post command transmits to the Communication-Manager of partner's system.

Event-Manager: It manages schema definition of events and their subscription, identifies whether subscription of an event is local or remote and transfer to corresponding Rule-Manager, which subscribes the event.

Rule-Manager: The Rule-Manager evaluates and executes ECA rules. It is implemented with basic trigger facility of underlined DBMS and contains Event-Instance-Table and Rule-Table.

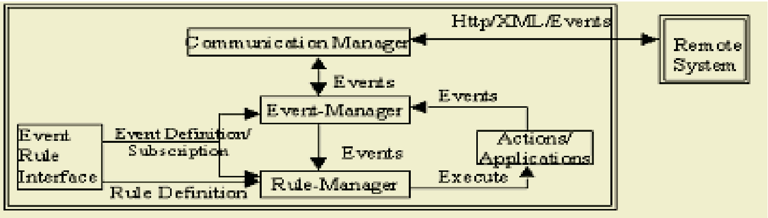


Fig. 4. Overall Architecture of Intelligent Active Functionality Component

Event-Rule-Interface: The Event-Rule-Interface is an interface for system administrators or programmers to define events and rules as well as to manage them, *i.e.*, search, delete, and update.

Actions/Applications: The actions/applications are internal if written in underlined DBMS's API or external if not.

Implementation and Evaluation of a Pilot System: To validate the timely collaboration mechanism and applicability of the IAFC, a pilot system has been implemented for a typical B2B scenario [7]. During design and implementation of the system, we considered practicability, interoperability with database, and platform independence. Therefore Java was chosen as an implementation language. Commercial DBMS Oracle 9i, Apache Tomcat Web server, Xerces2 Java parser were chosen.

For the application scenario, an internet shopping mall with its suppliers and shippers is used. The application is implemented with Pentium IV, Windows 2000 server, and Linux server. The implemented system has been tested to verify its correctness for all cases.

It should be noted that the timely collaboration mechanism is intended for emergency and asynchronous timely collaboration. Thus, it is complementary to workflow management system (WFMS) or business process management system (BPMS), which is intended for regular synchronous job process. Since the proposed IAFC is loosely coupled with other subsystems, it can be easily applied to other systems, which need active functionality.

6 Conclusion

In this paper collaboration among businesses in B2B E-commerce is analyzed and the need for timely collaboration is derived and classified in terms of inter-organizational contracts. A ECA rule based method for the timely collaboration and an IAFC to support it are proposed.

Since high-level ECA rule programming is supported by the IAFC, the collaboration among businesses and event-based immediate processing can be implemented independently to application logic. Thus, system administrators and programmers can easily program and maintain timely collaboration among businesses in B2B EC.

In order to extend our research future work includes support for various phases in B2B EC life cycle [8] and extension of collaboration concept to support client oriented ECA rules. Since Web Services and ebXML etc. are proposed as standard of EC systems recently, it is required to study the proposed IAFC to be integrated with them.

References

1. Nobuyuki Kanaya, et. al., "Distributed Workflow Management Systems for Electronic Commerce", proceedings of 4th International Enterprise Distributed Object Computing Conference(EDOC'00), IEEE 2000.
2. Norman W. Paton and Oscar Diaz, "Active Database Systems", Computing Surveys, ACM, 1999.
3. Brahim Medjahed, et al., "Business-to-business interactions: issues and enabling technologies", *VLDB Journal*, Springer-Verlag, April, 2003. pp.59-85.
4. Fabio Casati, S. Ceri, S. Paraboschi, and G. Pozzi, "Specification and Implementation of Exceptions in Workflow Management Systems", ACM Tr. on Database Systems, Vol. 24, No. 3, September 1999, pp.405-451.
5. Zongwei Luo, Amit Sheth, Krys Kochut, and Budak Arpinar, "Exception Handling for Conflict Resolution in Cross-Organizational Workflows“, Technical Report, LSDIS Lab, Computer Science, University of Georgia, April 10, 2002.
6. J. Meng, Stanley Y.W. Su, Herman Lam and A. Helal, "Achieving Dynamic Inter-Organizational Workflow management by Integrating Business processes, Events and Rules", IEEE HICSS-35'02, 2002.
7. DongWoo Lee, et al., "Active Functionality Component to Support Timely Collaboration among Businesses in B2B E-Commerce", J. of Computing Practice, KISS, vol.11 no. 2, 2005.
8. David Trastour, Claudio Bartolini, Chris Preist, "Semantic Web Support for the Business-to-Business E-Commerce Lifecycle", proceedings of www2002, 2002.

The Searching Methods of Mobile Agents in Telemedicine System Environments

Hyuncheol Jeong[1] and Inseob Song[2]

[1] 683-3 Shingchang-Dong, Kwangsan-Gu Kwangju, Dept. of Medical Information Engineering,
Kwangju Health College, Korea 506-701
hcjeong@kjhc.ac.kr
[2] Internet Research Center, GoldBank Communication Inc., Seoul, Korea
issong@www.goldbank.co.kr

Abstract. Telemedicine is an application area that combines the technology of computer and communication with medical service. Agent system can be applied to patient status watch and medical treatment system in medicine fields. A mobile agent autonomously recomputes its path according to the request of other agents or servers after it can report itself to the nearest host. Then, there is a time overhead. The location search of it can reduce time overhead for recomputing. In this paper, we propose an algorithm to reduce the time overhead and show its simulation through an analytical model. If mobile agent's migration path is changed while moving on the path, our proposed method searches precisely the location of the mobile agent with the changed path and reduces the time overhead.

1 Introduction

The development of information technology sets free a mass of data transmission and communication. Because information and communication system based on multimedia is applicable to medical system, telemedicine era comes true. Telemedicine is an application area to combine the technology of computer and communication, namely, various technologies of information and communication with medical service. Telemedicine is proper to unify medicine environment because telemedicine provides fast and accurately user with necessary medical information. Because telemedicine overcomes the local dependency of medical service, this system is applicable to remedy, diagnosis, tracing examination of surgical patient, chronic disease management, transmission of medical information, inspection of dosage. Telemedicine can improve the quality of medical service and reduce cost[1]. In medical field, agent system can be applied to patient status watch and medical treatment system. GUARDIAN[2] was designed for the purpose of watching severe surgical patient. It implemented the team consisted of experts in the internal parts of the system and was able to work cooperative watch through effective information exchange among team members. Agents to take charge of watching task are split into each function. These functions almost coincide with basic factor to construct agent. Agents to take charge of perception and action work in external world through sensor and effector. Inference agent processes task decision and control agent handles the overall system. Mobile agent(MA), autonomously movable process, executes user's task instead of user in application

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 728–734, 2005.

system. Many works [3,4,5,6] are in progress to research it. Because the distributed system environment to utilize mobile agent executes task by distributed cooperative processing among agents, it can improve system performance. So, communication among MA's is essential. Static agent(SA) operates in only system to occur it. Its position information easily searches in agent name server(ANS). ANS keeps the position information of every SA. But, a MA must trace the current location of another MA to communicate each other because a MA autonomously migrates to mobile agent system. In this paper, we propose an algorithm to trace the location of MA and analyze performance. For a MA to communicate with another MA in mobile agent system, whenever a MA, namely, communication target, changes its path, it notify changed path to another MA to desire communication.

2 Telemedicine System Using Mobile Agent

Telemedicine is a computer-based system to diagnose and remedy patients in remote sites. This system has some purposes to achieve as follows. First, the more healthful patients are, the less they go to hospital. Second, they are received treatment at home. Third, if they are very urgent, they go very fast to emergency room. Eventually, it is to increase other patient's opportunity cost. The structure of mobile agent in telemedicine system is in Fig. 1. Agent system is platform to manage the life cycle of agent such as creation, explanation, execution, abolition. To support MA, mobile agent system provides additional functions with general agent system like agent's mobility and communication, access restraint from unauthorized agent, etc. One or more mobile agent systems can be in host. Mobile agent system has one agent server, ANS, directory facilitator (DF) to manage directory within one domain, agent management server(AMS). Also, this system has a agent meeting point (AMP) and one or more place(i.e. agent execution environment). Each place contains one or more agents[5,6,7]. Agent is a software process autonomously to work instead of user. Place provides the execution environment of agent in system. AMP acts as the starting point and arrival point of MA. Also, it supports that agent utilizes host resource. DF has all agent names within a domain, service type, system address, message content and related attribute and it sends some information to agent request. DF is one within a domain. Agent information within another domain can be acquired by inquiring of DF. AMS manages the life cycle of agent. When agent migrates first, AMS registers it in DF. If it is not necessary, AMS delete it in DF. AMS is in all platforms to have SA one by one. It must know agent's path. Information exchange between agent systems works by cooperation among AMS's. ANS detects agent's name through DF and maintains the physical address of system to have that agent and provides information by agent request.

3 Communication Method Among Agents

Mobile agent can freely migrate among servers to provide service. Because it is under remote host control after migration, it can access resource despite disconnection between user and network. Agent system has the characteristics that it permits agent's migration among network nodes to support a MA[5]. In [7], communication methods among agents are classified into two parts.

3.1 Communication Among Static Agents

SA's are within the same agent system. All SA's within system write their position information in ANS. An agent inquires of ANS the location of another agent to communicate. Then, ANS to receive a query notifies agent's location. So, a agent to detect it can easily communicate with the agent to desire. Registration[7] takes less time than other method when it detects a mobile agent. But, if ANS has the migration path information of every MA and many agents want to communicate with a MA, overheads can be concentrated on ANS.

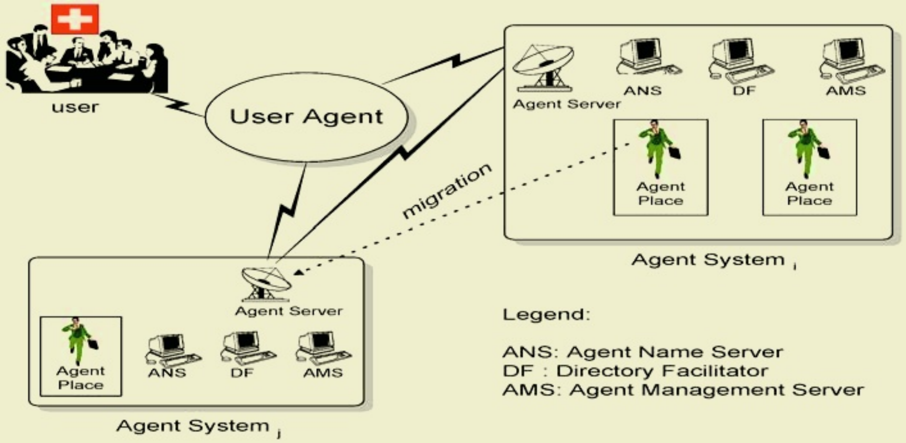


Fig. 1. Mobile agent in telemedicine system

3.2 Communication Among Static and Mobile Agents

An agent to be in agent system communicates with another mobile agent. There are four types. Broadcasting method sends messages to every agent server. So, this method searches an agent server where MA is. But, agent's migration velocity goes down because of network traffic among agent systems. Sequence search method sequentially sends messages to all agent servers and detects MA's location. This goes down system performance because the time of mobile agent's detection is late. Logging method detects agent's location by using the log information of each agent server. In the worst case, this must visit all agent servers. Thereafter, it detects MA's location. Also, much time is required to send position information to SA. Agent advertisement method maintains position information in ANS only when the status for user to define in agent is changed. If status is updated it has the same result as registration method whenever agent visits agent server.

4 Algorithm and Simulation

We describe necessary assumptions to propose algorithm. The algorithm is for mobile path and location. And we display simulation for that algorithm through analytical model.

4.1 Assumptions

To propose the algorithm that reduces the time overhead and effectively traces an agent's location, we utilize assumptions[7] as follows. Each agent system has one or more agent server to create agent and ANS. All agents have home server. They are created in home server. The ANS of each agent system maintains the address of agent server to have each agent location and the location of agent's home server. There are two types of agent. Search agent is in home server. When another agent requests, it detects the location of MA. Mobile mirror agent(MMA) also is in home server and observes ANS. If the identification of server to change the migration path of MA comes in ANS, it caches the identification and informs search agent of it. Each agent system has search agent and MMA to trace MA. ANS and MMA maintain the position information of MA.

4.2 Algorithm

MA informs ANS of the initial migration path for itself and move into the first agent server of initial migration path. MMA watches ANS. If ANS has the initial migration path, MMA detects the initial migration path of MA. If mobile agent's migration path is changed while MA moves on the agent server in the initial path order, the changed path is kept in current agent server's log. If the identification of agent server comes in ANS, MMA watches ANS and detects it. MA to change the path migrates to the first agent server of reconfigurated MA. When the path is changed again, these processing are repeated. So, ANS and MMA maintain the path information of MA. The algorithm to preserve the path information of MA is as follows.

```
Sept1: MA notifies the self-information of initial migration path to ANS.
Step2: MMA acquires this information from ANS.
Step3: MA sequentially migrates to the agent server of the initial path.
    Step3.1: If the mobile path is changed while moving,
        Step3.1.1: maintain the changed mobile path in the current agent server.
        Step3.1.2: notify the identification of agent server to keep the path to
                   ANS.
Step4: MMA in watching ANS acquires the identification of agent server if the identi-
    fication of agent server is maintained in ANS.
Step5: MA moves to the first agent of reconfigurated path.
Step6: If the path is changed again while moving, repeat from Step1 to Step6.
```

To communicate with MA, if an agent detects the position of MA, it acquires the location of home server of MA to desire from ANS. And it utilizes the search agent in home server and MMA. The algorithm to trace the location of MA is as follows.

```
Step1: MA acquires the location of home server from ANS.
Step2: MA requests the current location of MA from search agent in home server.
Step3: Search agent requests the mobile path information of MA to detect from
    MMA.
Step4: MMA informs search agent of the identification of agent server. Search agent
    has the initial or the changed path information of MA.
```

Step5: Search agent gets mobile path information from MMA.
- Step5.1: When the initial mobile path is received, search agent applies it to binary search.
- Step5.2: When the identification of server is received, search agent sends messages to server that corresponds to the identification and gets the changed path. Thereafter, apply the path to binary search.

Step6: Search agent utilizes binary search and traces the location of MA.
- Step6.1: If the mobile path of MA while the location is traced, MMA informs search agent of the identification of agent server in ANS.

Step7: Search agent stops binary search in progress.

Step8: Search agent sends messages to the server that corresponds to the identification of server to which MMA sends and gets the changed mobile path.

Step9: Again apply the changed mobile path to binary search and traces the location.

Step10: The mobile path of MA repeats from step5 to step9 while tracing.

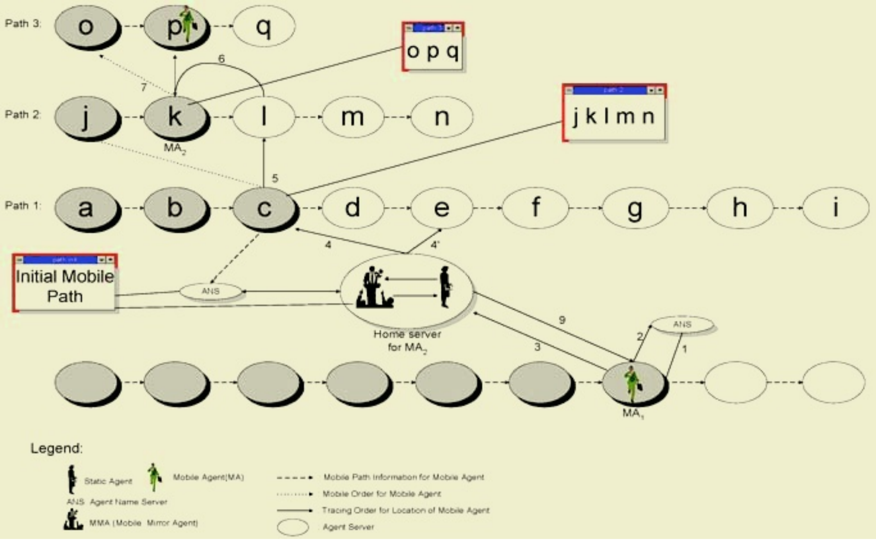


Fig. 2. MA's Location Searching

4.3 Simulation

We analyze the wasted time to trace the location of MA through analytical model. Parameters are as follows. O(T) is the wasted average time to trace the location of MA. m is the changed frequency for the path of MA. N is the number of agent server in the changed mobile path. T is the wasted time to send message from a agent server to another. C_i is the frequency of writing the identification of agent server to change the path in ANS while applying the algorithm for the i-th path. Whenever MA changes the path, MMA acquires the identification of server to have the changed path from ANS and detects the changed path. Therefore, the performance depends on the frequency for agent identification to be known in ANS. We consider the probability of frequency as follows.

$$P(X=C_i) = (\mu^{C_i} e^{-\mu}) / (C_i!) \quad (0 \leq P(X = C_i) \leq 1) \tag{1}$$

When the number of agent which is in mobile path is N_1 , N_2 , N_3 , …… , N_i, the wasted time equation to trace agent location is as follows.

$$O(T) = \log_2 N_m T + \Sigma_{i=1}^{m-1} P(X= C_i) \log_2 N_i T \tag{2}$$

To simulate, each parameter is assumed as follows. The number for agent to desire the communication with MA is 20. m is 10. N is 10 to 100. T is 5. ANS knows the server's location for mobile path to be changed. Each probability to be known the location is 0.1, 0.5, 0.9. Fig. 3 shows the relationship between the time to trace location and the number of agent server. While search agent executes location tracing for each mobile path, it acquires the identifier of agent server changed mobile path from MMA. Then the number of message is altered according to the frequency for MMA to acquire the identifier of agent server from ANS because of stopping location tracing for current mobile path. So, the number of message necessary to the location tracing of mobile agent is as follows. The number of agent to communicate with mobile agent is M and the number of agent in each mobile path is N_1, N_2, N_3, ..., N_i.

$$Msg = \log_2 N_m M + \Sigma_{i=1}^{m-1} P(X= C_i) \log_2 N_i M \tag{3}$$

Fig. 4. displays graph for the number of message necessary to location tracing and the number of agent server.

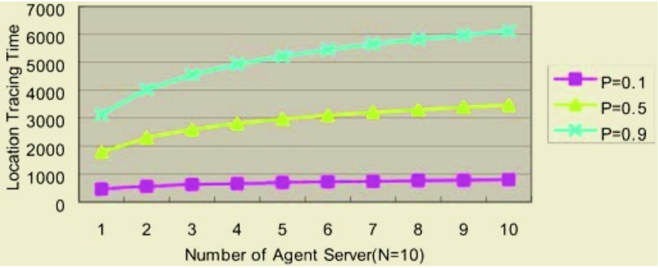


Fig. 3. Location tracing time based on probability

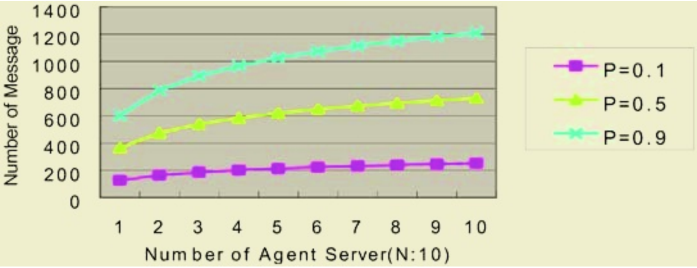


Fig. 4. Message necessary to location searching

5 Conclusion

In the existed methods, the wasted time for location tracing detracts the overall system performance when a mobile agent recomputes its path and migrates. In this paper, the migration path of mobile agent informs search agent of the changed identification of agent server by using search agent, MMA, and ANS. So, we proposed faster algorithm to trace the location of mobile agent and analyzed the time overhead through analytical model. In the future work, when network nodes cause fault, we will research an effective method to trace the agent location.

References

1. Joanne Kumekawa, Dena S. Puskin, Thomas Morris, *Telemedicine Report to Congress*, U.S. Department of Commerce (1997)
2. Barbara Hayes-Roth, Micheal Hewett, Richard Washington, Rattikorn Hewett, Adam Seiver," Distributed Intelligence within an Individual", In Les Gasser and Michael N. Huhns, editors, *Distributed Artificial Intelligence Volume II,* Pitman Publishing (1989)
3. Krishna A. Bharat, Luca Cardelli, Migratory Application, SRC Research Report, Digital Equipment Coporation (1996)
4. Harrison, C.G., Chess,D.M. and Kershenbaum, A., Mobile Agents: Are they a good idea?, IBM Research Report 19887, IBM Research Division (1995)
5. David Chess, Benjamin Grosof, Colin Harrison, David Levine, Colin Parris, Gene Tsudik, Itinerant Agents for Mobile Computing, Technical Report on IBM T.J. Watson Research Center (1996)
6. Jonathan Dale, A Mobile Agent Architecture for Distributed Information Management, A thesis submitted for the degree of Doctor of Philosophy (1997)
7. The Open Group, Mobile Agent Facility Specification, OMG TC Document, International Business Machines Corporation (1997)

Knowledge-Based RDF Specification for Ubiquitous Healthcare Services

Ji-Hong Kim[1], Byung-Hyun Ha[1,*], Wookey Lee[2],
Cheol Young Kim[3], Wonchang Hur[3], and Suk-Ho Kang[1]

[1] Department of Industrial Engineering, Seoul National University,
Seoul 151-742, Republic of Korea
valentine@ara.snu.ac.kr, pepper@netopia.snu.ac.kr
shkang@snu.ac.kr

[2] Department of Computer Engineering, Sungkyul University,
Anyang-8-dong, Manan-gu, Anyang, Korea
wook@sungkyul.edu

[3] CyberMed, Inc., 5F Won Bldg., Nonhyeon-dong, Gangnam-gu,
Seoul 135-814, Republic of Korea
kaster@cybermed.co.kr, hwc@cybermed.co.kr

Abstract. With ubiquitous computing the quality of human life can be improved by interoperation among various devices and services. These changes of computing paradigm are enabling the enterprises' legacy services to be automated and value-added all the more. We suggest a service framework and algorithms of provisioning healthcare services in a ubiquitous computing environment. In order to meet customers' need we translate the need into relevant goal and repeatedly refine the goal into sub-goals through commonsense knowledge until there are appropriate services for sub-goals and after, employ the services. The results of this research enable integration and interconnection of devices, applications, and functions within the healthcare services.

1 Introduction

With ubiquitous computing the quality of human life can be improved by interoperation among various devices and services [8]. To realize the new computing paradigm, various efforts are being conducted. By the help of these efforts, simple and direct needs of users, e.g. turning on and off lights or playing music automatically according to a user's preference, can be easily accomplished. However when the needs are indirect and the tasks that fulfill the needs are complex, we need the ingenious mechanism that coordinates a variety of objects including the users themselves [4].

In order to meet the user's requirements at high level we need an intelligent framework which seamlessly integrates devices and applications in a ubiquitous computing environment. For the first step, it is indispensable to represent unit functionalities provided by the ubiquitous computing environment as the concept of services. *Services* in a ubiquitous computing environment are defined as atomic functionalities that are available to users. Modeling services apart from the devices which provide unit functions has the advantage of separating actual implementations from the physical devices or applications. As a result, necessary services can be accessed

[*] Corresponding author. Tel.:+82-2-880-7180, fax.+82-2-889-8560

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 735–741, 2005.

transparently and flexibly, therefore it becomes easy to build plans that satisfy users' needs.

When services are modeled for the functionalities of devices, the services need to be combined dynamically according to the situation using knowledge base in other to achieve high-level requirements of users. The reasons for dynamic composition of services can be stated with three perspectives of users, devices, and domains.

First, at user perspective, the needs of users in reality seem too diverse to define all possible schemes of services in advance. If we define every service scheme statically, it will be have limited scalability. This is because almost all services need to be redefined, even when the details of user's needs change slightly. At the device perspective, the availability of devices and applications making up ubiquitous environment changes according to time and place. Therefore the services that they can provide are impossible to predict and it is not applicable to assume fixed conditions. Lastly, in order to apply one successful system in a certain domain to anther domain, the system requires to be designed in a generic form. If a system is implemented using generic rules and domain-specific knowledge to meet the needs of users, the system can be easily applicable in another domain only after modifying the domain-specific knowledge. As a result, it is important to compose services dynamically based on predefined knowledge to build successful ubiquitous computing.

In this research we suggest a service framework and algorithms of provisioning healthcare services in a ubiquitous computing environment. The framework is composed of atomic services, user's goals, goal-based agents, dynamic service composition algorithm, and knowledge bases. Currently most researches of composing services dynamically depend on the conventional planning methods of Artificial Intelligence domain. This approach stands on the Mikrokosmos ontology [6] which enables querying and answering based on shared knowledge base and on the HowNet [2] for causality inference based on the set of lexical knowledge bases. All these show that the network of concepts is used for extracting more refined knowledge.

2 Goal-Based Service Provision Agent

Mike has registered at uHealthNet Inc. for diabetic management services. Just like every other day, Mike went to the bathroom, sat on the toilet. Mike's health information is forwarded to uHealthNet. uHealthNet's doctor realizes that Mike has hyperglycemia. The doctor can now tell the agent to notify Mike that he needs to take the insulin shot within 2 hours. Mike's agent then checks the insulin supply at Mike's home and realizes that the supply is insufficient. One this is noticed, a nearest pharmacy that has the supply of insulin is located and an appropriate delivery service is also selected. After "Pharmacy A" has been chosen, prescription is sent to "A."

Upon reviewing Mike's schedule, the agent realizes that Mike needs to be at work within an hour. Now, Mike's agent sends a message to the delivery service notifying them to deliver the insulin from "Pharmacy A" to Mike's office after 1 hour. Now, Mike gets up at toilet. And the "Smart Mirror" say, "You are currently diagnosed as hyperglycemia and therefore need to receive and insulin shot within 2 hours. Current supply of insulin is insufficient so your new supply will arrive at the same time you arrive at your office – end of message." Mike views this message, finishes the rest of his morning activities and goes off to work. As Mike arrives at this office, the insulin is delivered simultaneously and Mike takes the insulin shot.

Through this scenario, we can derive needs considering the important aspects of goal management. Namely, there is a time gap between the achievement of the goal and the execution of service. The agent needs to be aware of this time gap and can't keep determining new services until the goal is completed. If it does so, conflicts will occur as new goals are continuously made. More importantly, Mike is supposed to receive the service after one hour when he arrives at his office. But, there is a chance that this service will not be completed. If it does agent need to reconfigure service for goal. So we notice that defining, managing and maintaining the goals in order to provide the necessary services are the main points in ubiquitous computing environment. With this in mind, we describe the high level logic for defining, managing, and maintaining goals. When the goal is received by the agent, it needs to realize the specific state of the goal whether it is *incomplete*, *pending*, *completed*, or *terminated*.

In the state of incompletion, a plan needs to be determined in order to complete the goal. The pending state is when the service is actually in motion to complete the goal. In this state, service is continually in motion until the goal is completed or an outside event interferes with the current state. While the goal is in a pending state it can end in terminated state before reaching a completed or incomplete state. The goal in its pending state alters into incomplete state again, either when achieving the goal ends in failure due to an exception, or when new opportunities appear to attain the goal in a more appropriate way. And whenever a goal turns into an incomplete state, an agent builds plans to complete the goal as soon as possible. Likewise, when an agent decides the goal in the complete state is invalid because of outside events, the goal turns into an incomplete state. When a one time applicable goal is completed or when a certain goal is no longer preferable by a user any more, the goal is altered into terminated state. Fig. 1 shows the approach of our Goal-Based Agent.

```
function GOAL-BASED-AGENT(event)
        static:  KB, a knowledge base
                 goal_list, a list of goals
                 new_goal ← MAKE_GOAL(event, KB)
        if new_goal is not null then
                 append new_goal to goal_list
        for each goal in goal_list
                 if filter of goal approves event then
                         EXECUTE-AND-UPDATE-GOAL-STATE(goal, event)
                         if state of goal is incomplete then
                             plan ← KNOWLEDGE-BASED-PLANNER(goal, KB)
                             set plan for goal
                         else if state of goal is terminated then
                             remove goal from goal_list
        return
```

Fig. 1. An agent that performs goal-based planning. It manages goals based on the assumption that each goal is independent to each other. Knowledge-Based Planner is explained in the next

3 Knowledge-Based Service Composition Algorithm

To compose services dynamically conventional approach directly connect user's goal to unit service's output. So the goal is achieved by using only the combination of unit services. This approach has many limits as we refer to above. In order to overcome limits, we will fill the missing link between services using approach that refine the goal. We grasp the meaning of a goal that is complex, various and user-centric by reasoning through knowledge-base, and refine the goal using methods that refine the goal. Because of this, the meaning of goal can be more precise and explicit. Through this, we can achieve to provide a more flexible service by using an appropriate service.

3.1 Knowledge Base for Service Composition

3.1.1 Semantic Part

The semantic part is used for interoperability as vocabulary for communication between agents and basis of knowledge base. The semantic part defines, represents, and describes concepts that have semantic like object, relation between object, object's state, object's operation that has semantic and etc. Because of this, semantic part enables that agents can reason by jointly holding the meaning. Through the same reason, agents can grasp hierarchy, similarity between concepts, and agent can grasp roles, characters, and attributes of concepts. The semantic part basically has a triple form that consists of subject, predicate and object. Subject and object is represented by concept, and predicate has a role to connect between the concepts, and is represent by property and hierarchy like super-class and sub-class.

3.1.2 Logic Part

Logic part defines and represents principle, rule and knowledge between facts. Logic part enables to make user's goal and sub-goal that are newly created into a more precise and explicit goal. More precise and explicit goal enables to search an appropriate service easily and accurately.

Logic part also consists of triple, but it is a little different from the semantic part. Namely, subject and object of the logic part consist of another triple, and predicates of the logic part consist of two kinds of properties. First is the precondition property that means that a triple is surely operated as another triple's precondition. Next is the enabling property that means that a triple enables to convert into another triple.

If a triple connects with many triples, properties have the form that is "AND" or "OR" split. A triple that is connected by "OR" split has a priority. As a priority exists, it enables to not check all situations but selecting one that has high priority. Because of this, goal is grasped and we can refine it. And a priority helps not only reduction of computational complexity but also expands to a more rational goal. Fig. 2 is partial logic part in scenario of Section 2

3.1.3 Service Part

Service part defines and represents unit service that is provided to the user. Unit service contains functions of device, software, and network, and contains the flow of information between functions [7]. Service part has four properties. They are input, output, precondition and effect.

(∀x: User) (∀y: Action)

$$\mathrm{Act}(x, y) \xrightarrow{\mathit{PRECONDITION}} \mathrm{Motivated}(x, y)$$

“User X act an action Y” has precondition “X is motivated about y”

(∀x: User) (∀y: food)

$$\mathrm{Eat}(x, y) \begin{array}{l} \xrightarrow{\mathit{PRECONDITION}} \mathrm{Have}(x, y) \\ \qquad\qquad\quad \textbf{\textit{AND}} \\ \xrightarrow{\mathit{PRECONDITION}} \mathrm{exist}(y) \end{array}$$

“User X eat a food Y” has precondition “X have Y” and “Y exist”

(∀x: User) (∀y: material)

$$\mathrm{Have}(x, y) \xrightarrow{\mathit{PRECONDITION}} \mathrm{Close}(x, y)$$

“User X have a material Y” has precondition “Y is close to X”

(∀x: User) (∀y: material)

$$(*) \quad \mathrm{Close}(x, y) \begin{array}{l} \xrightarrow{\mathit{PRECONDITION}} \textbf{\textit{1}}\ \mathrm{go}(x, y) \\ \qquad\qquad\qquad\qquad \textbf{\textit{OR}} \\ \xrightarrow{\mathit{PRECONDITION}} \textbf{\textit{2}}\ \mathrm{go}(y, x) \end{array}$$

“User X is close to material Y” has precondition “X go to Y” or “Y go to X”, “Y go to X” has high priority

(∀y: material) (∀x: User)

$$\mathrm{Go}(y, x) \xrightarrow{\mathit{ENABLING}} \mathrm{Supply}(x, y)$$

“Material Y go to User X” convert to “Y is supplied to X”

Fig. 2. Partial logic part. (*) is an example of having two precondition properties in OR split. The numbers at the right sides of precondition arrows mean the priorities of the preconditions

3.2 Service Composition

In order to refine and manage goal accurately for service composition, knowledge-base will be provided perfectly. We assume this, and we present an algorithm and a semantic similarity measure for service composition in this section.

3.2.1 Service Composition Algorithm

In this section, we explain an algorithm for service composition in detail. User’s goal is given as a triple. After acquiring basic information of each of triple’s elements through searching in knowledge-base, service composition is started.

Step 1: Discovering Service for Goal. If a goal is given, it first searches services whose index matches the predicate or object of the goal. Then it checks all similarities by matching each combination of the searched result of the predicate, object and their synonyms. We explain the semantic similarity measure in detail at step 2. If semantic similarity is greater than the threshold, that particular service is chosen. Since determining specific threshold is beyond the scope of this research, we did not mention.

Step 2: Discovering Logic for Goal and Goal Refinement. At step 1, if it cannot search appropriate services according to the goal, first it searches logic whose subject matches to the predicate or object of the goal. Then it checks the semantic similarity by matching the subjects of logic and the goal. If the semantic similarity is greater than the threshold, this particular logic is used. If the logic has a precondition property only or enabling property only, the object of the logic becomes the new sub-goal, then it goes to step 1. If logic simultaneously has both a precondition property and an enabling property, the first object that connects to the enabling property becomes the new sub-goal, and then it goes to step 1. When the same logic is selected once more and if this sub-goal is achieved, the object that connects to precondition property becomes the new sub-goal and again it goes to step 1. And if searched logic does not match the goal, the subject of the logic is then queried to present state. If the answer is yes, the object of logic obviously becomes the new sub-goal.

Step 3: Selecting Service. If services that have the same function are searched at step 1, the service that has the minimum cost is selected. There are many kinds of costs; those are distance from the user, cost for using the service, response time of the service, reliability of the service and reputation of service. We select the service among many services, only considering the distance from the user.

Step 4: Service Execution. If one service is selected at step 3, the input and precondition of the service will be checked. If all input and precondition is satisfied with the output or the effect of other services that are already executed or at present state, service is executed. Otherwise, input and precondition of the service are the new sub-goal. Then it goes to step 1.

3.2.2 Semantic Distance Measure

We obtain the value of similarity between services as the weighted sum of similarities between entities – input, output, precondition and effect that consists of service.

$$\begin{aligned} S(A,B) &= W_{in} \times S_{in}(A_{in},B_{in}) + W_{out} \times S_{out}(A_{out},B_{out}) + W_{pre} \times S_{pre}(A_{pre},B_{pre}) \\ &+ W_{eff} \times S_{eff}(A_{eff},B_{eff}) + W_{index} \times S_{index}(A_{index},B_{index}) \end{aligned} \quad (1)$$

$S(A, B)$ represents the similarity between service A and B. W_i represents the weight of each entity. $S_i(A_i, B_i)$ represents the similarity between each entity of services.

$$S_x(A_x,B_x) = 1 - \underset{\substack{Z_0,Z_1,\ldots,Z_{r+1}\in\Sigma \\ Z_0=X,Z_{r+1}=Y}}{MIN}(1-\prod_{i=0}^{r}(1-d'(Z_i,Z_{i+1}))) \quad x\in\{in,out,pre,eff\} \quad (2)$$

We obtain the value of the similarity of each entity, using the semantic distance measure [3]. In Equation 2, A_x and B_x represent entities, $S_x(A_x, B_x)$ represents semantic similarity between A_x and B_x, and $d'(Z_i, Z_{i+1})$ represents direct distance between the entities. We obtain value of $d'(Z_i, Z_{i+1})$ using the entity matching algorithm in [1].

4 Conclusions and Future Works

In order to verify the methods that are represented in our research, we embody the prototype system. We construct domain ontology that is related to diabetes and minimum common sense ontology for service composition using the RDF Schema, and knowledge-base is represented according to this RDF Schema as RDF. We simulated company's back-end function to provide healthcare service through the use of simple expert systems. We also simulated devices and agents to interact with the user using

Visual C++ in MS Windows environment. Knowledge-based reasoning that is based in ontology is executed using SWI-Prolog engine through Horn logic, and we connected virtual components to the reasoning engine using C language interface of the SWI-Prolog. Components communicate with each other using RDF messages.

Experiments with limited number of services in virtual environment show that the approach of our research is feasible. We expect that our research will provide a foundation that realizes these efforts faster and provide better quality service.

However in order to cope with various services and unpredictable scenarios, enormous common sense knowledge will be needed. This is not easy similar to the limitation of the Cyc Ontology Project [5]. In order to rationally cope with the actual situation, decisions in uncertain situations and considerations of the feasibility of devices and services are also needed.

References

1. Aversano, L., Canfora, G., Ciampi, A.: An algorithm for Web service discovery through their composition. In: Proceedings of the IEEE international Conference on Web services (ICWS'04), California (2004) 332-339
2. Choi, K.-S., Kim, J.-H., Miyazaki, M., Goto, J., Kim, Y.-B.: Question-Answering Based on Virtually Integrated Lexical Knowledge Base. In: Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages, Sapporo (2003) 168-175
3. Cooper, M.C.: Semantic Distance Measures. Computational Intelligence. 16 (2000) 79-94
4. Fujii, K., Suda, T.: Dynamic Service Composition Using Semantic Information. In: 2nd International Conference on Service Oriented Computing, New York (2004)
5. Lenat, D. B.: CYC: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM, 38 (1995) 33-38
6. Shin, H., Koehler, S.: A Knowledge-Based Fact Database: Acquisition to Application. In: Proceedings of the International Conference (KBCS2000), Mumbai (2000)
7. Sycara, K., Paolucci, M., Ankolekar, A., and Srinivasan, N.: Automated discovery, interaction and composition of Semantic Web services. Web Semantics: Science, Services and Agents on the World Wide Web. 1 (2003) 27-46
8. Weiser, M.: The Computer for 21st Century. Scientific American. 265 (1991) 94-104

A Time Judgement System Based on an Association Mechanism

Seiji Tsuchiya[1,2], Hirokazu Watabe[1], and Tsukasa Kawaoka[1]

[1] Dept. of Knowledge Engineering & Computer Sciences, Doshisha University
Kyo-Tanabe, Kyoto, 610-0394, Japan
[2] Human Ecology Research Center, R&D H.Q., Sanyo Electric Co,. Ltd.
Hirakata, Osaka, 573-8534, Japan

Abstract. Common sense and judgement ability, in the same way as humans, are necessary to realize a computer that communicates directly with humans. Such a thing needs the ability to recall a concept from a particular word and to associate the concept with others. A Time Judgement System that can understand everyday time expressions is important for natural communication, and the system was based on the aforesaid Association Mechanism. The purpose of this research is to construct the system that can treat regular time expressions with adaptability to unknown expressions. The resultant Time Judgement System achieves a correct response rate of 75.9% with accuracy of 85.8%, comparable to that of human subjects.

1 Introduction

Our group is conducting research on the development of an intelligent robot that can converse naturally with people. Here, "intelligent" refers to the ability to use common sense, in the same way as humans, to understand and make judgements and then to respond and act. The ability to recall a concept from a particular word and to link that concept to various other related concepts plays an important part in achieving this objective. A method for the construction of a mechanism that allows one concept to be associated with various other concepts has already been developed [1][2], and a method for calculating the Degree of Association has also been proposed [3]. Using this Association Mechanism, a system for judging sentiments called to mind by humans given a certain noun on a common-sense basis using this association has been developed recently [4][5]. In this paper, a method for realizing judgements relating to time is proposed as one type of common-sense judgement made by humans.

This research involves the construction of a flexible mechanism that allows a robot to accommodate expressions with which it is unacquainted by drawing on everyday time expressions. This research also addresses the question of how a small amount of knowledge can be used with greater diversity from a time standpoint, such as by guessing time from wording, without requiring combination patterns of substantive and declinable parts.

2 Time Judgement System

Figure 1 shows the structure of the Time Judgement System, which consists of a knowledge base containing words that express time, and an Unknown-Word Process-

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 742–748, 2005.

ing process that takes words that are not known to the system and establishes a relationship between the unknown words and known words, allowing unknown words to be handled as known words. The Unknown-Word Processing generates word associations using a Concept Base [1][2], which is a large-scale database formulated automatically and manually by taking words from multiple electronic dictionaries. An algorithm is used to calculate the Degree of Association and evaluate the relationship between one word and another [3], labeled here as the Association Mechanism.

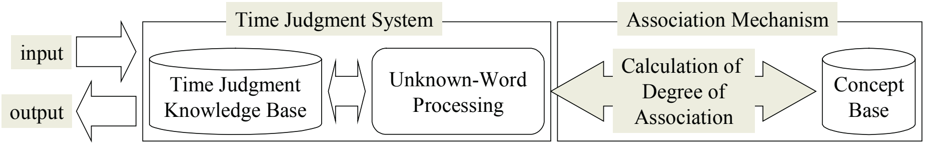


Fig. 1. Structure of Time Judgement System

This research considered the following expressions pertaining to time:

(1) Substantive words and phrases formed by combining substantive words (e.g., “morning” and “tomorrow morning”).

(2) Phrases that combine substantive and declinable words (e.g., “the sun rises” and “leaves are red”).

These were then handled as linguistic constructions that allow people to judge the time from a single phrase. In the examples given above, an association can be made between “the sun rises” and “morning”, while “colored leaves” can be envisioned from “leaves are red”, and from that, it can be judged that the season is autumn. In other words, the processing does not handle linguistic constructions from which even human cannot judge time, such as the phrase “commodity prices rise”, or from which people cannot judge time simply from that phrase alone, such as “the day of the incident”.

The understanding of (1) is called Time-Word Understanding, and the process of elucidating (1) from (2) is called Time-Word Generation. An understanding of time is defined as the ability to make a judgement as to whether or not a certain phrase is a time-related phrase, and additionally, if the phrase is time-related, to produce the time that is evoked by the phrase.

Times presented as a result of the judgement are specific numbers (for example, Christmas = December 25, afternoon = 12:00 to 23:59), or the nine words that are registered as the absolute time words that are Explicit Time Words in the Time Judgement Knowledge Base as described in Section 3: “spring”, “rainy season”, “summer”, “autumn”, “winter”, “morning”, “noon”, “evening”, and “night”. These nine words are referred to as Representative Time Words.

The specific flow of Time Judgement is as follows. When a word is input, the Time Judgement Knowledge Base is referenced, and a judgement is made as to whether the word is known or unknown. If the word is a known word, the specific numeral registered in the Time Judgement Knowledge Base is output. If the word is unknown, the Unknown-Word Processing described in Section 5 is carried out and a Representative Word is output. If a phrase was input, the processing is carried out in the same way for each word, and the time for that phrase is judged based on the majority decision for the time expressions that were judged.

3 Time Judgement Knowledge Base

The Time Judgement Knowledge Base is generally divided into explicit time words and suggestive Time Words. For the present Time Judgement System, the knowledge base contained 565 Time Words, which is thought to be the minimum vocabulary necessary. Explicit Words are words that in themselves express time. A total of 378 words are registered in the knowledge base, divided into nine classifications. Suggestive Time Words are words that in themselves do not indicate time, but which suggestively evoke associations, such as elucidating “winter” from “ski”. There are 187 words registered in the knowledge base as suggestive time words.

4 Concept Base and Degree of Association Algorithm

The Association Mechanism consists of the Concept Base and the Degree of Association Algorithm. The Concept Base generates semantics from a certain word, and the Degree of Association Algorithm uses the results of the semantics expansion to express the relationship between one word and another as a numeric value.

4.1 Concept Base [1][2]

The Concept Base is a large-scale database that is constructed both manually and automatically using words from multiple electronic dictionaries as concepts and independent words in the explanations under the entry words as concept attributes. In the present research, a Concept Base containing approximately 90,000 concepts was used, in which auto-refining processing was carried out after the base had been manually constructed. In this processing, attributes considered inappropriate from the standpoint of human sensibility were deleted and necessary attributes were added.

train	train, 0.36	locomotive, 0.21	railroad, 0.10		a_i, w_i	Primary Attributes
	train, 0.36	locomotive, 0.21	railroad, 0.10	...	a_{i1}, w_{i1}	Secondary Attributes
	locomotive, 0.21	streetcar, 0.23	subway, 0.25	...	a_{i2}, w_{i2}	
	:	:	:	:	:	
	a_{1j}, w_{1j}	a_{2j}, w_{2j}	a_{3j}, w_{3j}	...	a_{ij}, w_{ij}	

Fig. 2. Example demonstrating the Concept “train” expanded as far as Secondary Attributes

In the Concept Base, Concept A is expressed by Attributes a_i indicating the features and meaning of the concept in relation to a Weight w_i denoting how important an Attribute a_i is in expressing the meaning of Concept A. Assuming that the number of attributes of Concept A is N, Concept A is expressed as indicated below. Here, the Attributes a_i are called Primary Attributes of Concept A.

$$A = \{(a_1, w_1), (a_2, w_2), ..., (a_N, w_N)\}$$

Because the primary Attributes a_i of Concept A are taken as the concepts defined in the Concept Base, attributes can be similarly elucidated from a_i. The Attributes a_{ij} of a_i are called the Secondary Attributes of Concept A. Figure 2 shows the elements of the Concept “train” expanded as far as the Secondary Attributes.

4.2 Degree of Assosiation Algorithm [3]

For Concepts A and B with Primary Attributes a_i and b_i and Weights u_i and v_j, if the numbers of attributes are L and M, respectively ($L \leq M$), the concepts can be expressed as follows:

$$A = \{(a_1, u_1), (a_2, u_2), ..., (a_L, u_L)\}$$
$$B = \{(b_1, v_1), (b_2, v_2), ..., (b_M, b_M)\}$$

The Degree of Identity $I\,(A, B)$ between Concepts A and B is defined as follows (the sum of the weights of the various concepts is normalized to 1):

$$I(A,B) = \sum_{a_i = b_i} \min(u_i, v_j)$$

The Degree of Association is calculated by calculating the Degree of Identity for all of the targeted Primary Attribute combinations and then determining the correspondence between Primary Attributes. Specifically, priority is given to determining the correspondence between matching Primary Attributes. For Primary Attributes that do not match, the correspondence between Primary Attributes is determined so as to maximize the total degree of matching. Using the degree of matching, it is possible to give consideration to the Degree of Association even for Primary Attributes that do not match perfectly.

When the correspondences are thus determined, the Degree of Association $R\,(A, B)$ between Concepts A and B is as follows:

$$R(A,B) = \sum_{i=1}^{L} I(a_i, b_{xi})(u_i + v_{xi}) \times \{\min(u_i, v_{xi}) / \max(u_i, v_{xi})\} / 2$$

In other words, the Degree of Association is proportional to the Degree of Identity of the corresponding Primary Attributes, and the average of the weights of those attributes and the weight ratios.

5 Unknown-Word Processing Technique

If an unknown word is synonymous with a known word or has an extremely strong association with it, the Degree of Association is strong in terms of both meaning and time representation. Moreover, it is extremely difficult, and unrealistic, to create knowledge for all of the words that pertain to a certain word, and to store the knowledge in a database. Thus, a very small number of representative words that can efficiently express time-related concepts are selected, and those words are stored in the Time Judgement Knowledge Base. The Degree of Association between the known words stored in the base and a given unknown word is then evaluated, and known words that have a strong Degree of Association with the unknown word are returned. This makes it possible to handle unknown words with equivalent power as known words. This processing is called Unknown-Word Processing, as described below.

5.1 Substitution Processing of Words with the Highest Degree of Association

This is an Unknown-Word Processing Technique in which the Degree of Association Algorithm is used and known words with an extremely strong Degree of Association with the unknown word are elucidated.

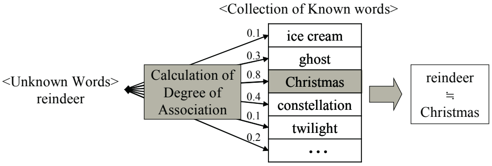


Fig. 3. Specific example of substitution processing of words with the highest Degree of Association

(1) The Degree of Associations between the unknown word *X* and all of the known words are calculated.

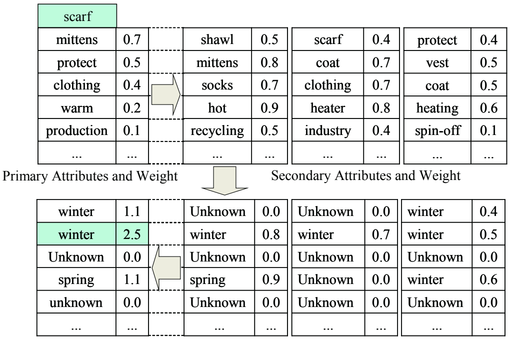
(2) The known word with the largest Degree of Association with the unknown word *X* is returned for the unknown word, and the Representative Time Word corresponding to that word is taken as the time expressed by the unknown word *X*. However, a threshold *Thr* is provided, and if the largest Degree of Association is at or below the threshold *Thr*, the known word is not returned for the unknown word *X*, and it is judged that the word is unrelated to time. The value of 0.8 was determined through testing to be most effective for the threshold *Thr*. Figure 3 shows a specific example of the substitution processing of words based on the highest Degree of Association.

5.2 Majority Decision Unknown-Word Processing with Secondary Threshold

This is an Unknown-Word Processing Technique based on the attributes and weight of the concept.

(1) The Primary Attributes of an unknown word *X* are obtained from the Concept Base.

(2) A search is carried out to determine whether any of the Secondary Attributes of the unknown word *X*, meaning the Primary Attributes of X_n (X_{n1}, …, X_{nm}), are known words that exist in the Time Judgement Knowledge Base. If a Secondary Attribute is a known word, the Representative Time Word associated with that word is returned for the attribute.


scarf
Primary Attributes and Weight
Secondary Attributes and Weight

mittens	0.7	shawl	0.5	scarf	0.4	protect	0.4
protect	0.5	mittens	0.8	coat	0.7	vest	0.5
clothing	0.4	socks	0.7	clothing	0.7	coat	0.5
warm	0.2	hot	0.9	heater	0.8	heating	0.6
production	0.1	recycling	0.5	industry	0.4	spin-off	0.1
...	...	...	...	...	...	...	...

winter	1.1	Unknown	0.0	Unknown	0.0	winter	0.4
winter	2.5	winter	0.8	winter	0.7	winter	0.5
Unknown	0.0	Unknown	0.0	Unknown	0.0	Unknown	0.0
spring	1.1	spring	0.9	Unknown	0.0	winter	0.6
unknown	0.0	Unknown	0.0	Unknown	0.0	Unknown	0.0
...	...	...	...	...	...	...	...

Fig. 4. Specific example of majority decision Unknown-Word Processing with secondary threshold

(3) From among X_{n1}, …, X_{nm}, the weight of X_n for the attribute returning a Representative Time Word is taken as the index. The indices for each of the Representative Time Words are added, and the Representative Time Word with the largest index exceeding the lower limit *Ths* is returned for X_n. If multiple Representative Time Words exist for the same index, the word with the largest Degree of Association is returned. If X_{n1}, …, X_{nm} are all unknown words, X_n is taken to be an unknown word.

(4) If a Representative Time Word was returned for X_n, the weight W_n for X_n pertaining to unknown word *X* is added to the index of that Representative Time Word. This is repeated for all of the Primary Attributes.

(5) The Representative Time Word with the largest index exceeding *Thv* is returned for the unknown word *X*. If multiple Representative Time Words exist for the same index, the word with the largest Degree of Association is returned.

Figure 4 shows a specific example of the majority decision Unknown-Word Processing method with a secondary threshold. Values of 0.025 and 0.1 were determined for *Ths* and *Thv* through preliminary testing.

5.3 Two-Stage Unknown-Word Processing Technique and Evaluation

The test data presented below were used to evaluate the performance of the two-stage Unknown-Word Processing Technique. The test data include words both related and unrelated to time, compiled in advance based on questionnaire results. Seven test subjects were asked to judge whether the various data were related to time, and only those words for which five or more subjects returned the same judgement were used in the test data as the results of judgements made by humans.

Table 1. Evaluation of the performance of the two-stage Unknown-Word Processing Technique

	Words related to time	Words unrelated to time	total
Correct response rate	75.4%	97.6%	86.5%
Accuracy	90.5%	97.6%	94.1%

– Words related to time (289 words): Words used daily from a list compiled from questionnaires and season words used in haikai poetry to express time.

– Words unrelated to time (250 words): Words compiled from questionnaires.

The results were evaluated by dividing the Time-Word Judgement results for the test data into three classifications: "correct response", "incorrect response", and "no response". The rate of correct responses and the accuracy were defined by

Correct response rate = No. of correct responses / (no. of correct responses + no. of incorrect responses + no. of no responses)

Accuracy = No. of correct responses / (no. of correct responses + no. of incorrect responses)

The evaluation results for the test data are presented in Table 1.

6 Evaluation of Time Judgement System

The Time Judgement System was evaluated using the test data listed below.

Group A: Explicit time expressions in newspaper articles (285 expressions)
Group B: Phrases related to time (256 phrases)

Group C: Words related to time (289 words)
Group D: Words unrelated to time (250 words)

The evaluation results for each of the test data groups are presented in Table 2.

An average of 75.9% correct responses was obtained in this test, with an average accuracy of 85.8%. These rates are comparable to the judgements made by humans, indicating that the Time Judgement System with two-stage Unknown-Word Processing is effective.

Table 2. Evaluation of various data groups

Test data groups	No. of correct responses	No. of incorrect responses	No. of no responses	Average of correct responses	Accuracy
Group A	235	50	-	82.5%	82.5%
Group B	123	57	76	48.0%	68.3%
Group C	218	23	48	75.4%	90.5%
Group D	244	6	-	97.6%	97.6%

7 Conclusions

A method for realizing judgements pertaining to time based on concept association was presented. Time Judgement is one type of common-sense judgements made by humans. Focusing on everyday time expressions, with basic common-sense knowledge provided in advance, the proposed method can accommodate numerous unknown expressions for which no knowledge exists.

The Time Judgement System achieved an average correct response rate of 75.9%, with an average accuracy of approximately 85.8%. These rates are comparable to those for judgements made by humans, demonstrating that the Time Judgement System is effective.

Acknowledgements

This work was supported with the Aid of Doshisha University's Research Promotion Fund.

References

1. Hirose, T., Watabe, H. and Kawaoka, T.: "Automatic Refinement Method of Concept-base Considering the Rule between Concepts and Frequency of Appearance as an Attribute", Technical Report of the Institute of Electronics, Information and Communication Engineers, NLC2001-93, pp.109-116 (2002)
2. Kojima, K., Watabe, H. and Kawaoka, T.: "A Method of a Concept-base Construction for an Association System: Deciding Attribute Weights Based on the Degree of Attribute Reliability", Journal of Natural Language Processing, Vol.9, No.5, pp.93-110 (2002)
3. Watabe, H. and Kawaoka, T.: "Measuring Degree of Association between Concepts for Commonsense Judgements", Journal of Natural Language Processing, Vol.8, No.2, pp.39-54 (2001)
4. Horiguchi, A., Tsuchiya, S., Kojima, K., Watabe, H. and Kawaoka, T.: "Constructing a Sensuous Judgement System Based on Conceptual Processing", Computational Linguistics and Intelligent Text Processing (Proc. of CICLing-2002), Springer-Verlag, pp.86-95 (2002)
5. Watabe, H., Horiguchi, A. and Kawaoka, T.: "A Sense Retrieving Method from a Noun for the Commonsense Feeling Judgement System", Journal of Artificial Intelligence, Vol.19, No.2, pp.73-82 (2004)

Response-Driven Web-Based Assessment System

Sylvia Encheva[1] and Sharil Tumin[2]

[1] Stord/Haugesund University College, Bjørnsonsg. 45, 5528 Haugesund, Norway
sbe@hsh.no

[2] University of Bergen, IT-Dept., P.O. Box 7800, 5020 Bergen, Norway
edpst@it.uib.no

Abstract. This paper focuses on a response-driven Web-based assessment system enhancing learning. The system responds to students' needs as they progress through the system. A structured environment, suitable for producing automatic help through logic-based and qualitative reasoning mechanisms guiding students in their navigation, is provided.
To decide when to recommend a student to solve another problem, it is necessary to assess the student's knowledge and the level of confidence she has in her answers. This is achieved in an automatic way allowing students to work without a human tutor.

1 Introduction

Gathering and interpreting information about student learning is an important part of any educational process. An intelligent Web-based assessment system uses such information to diagnose the strengths and weaknesses of student learning and then provides further guidance. If the results of a test demonstrate that a student does not understand a concept, the system provides supplementary material to re-teach the concept.

Making high-quality assessment tests supporting learning requires an understanding of how students learn [6] and what is important for them to learn. There is a need for tests assessing the level of students' ability to use knowledge in an interrelated way when analyzing and solving authentic problems. Questions in such tests will ask students to compare different methods, draw together several concepts, evaluate consequences, and handle new situations. Research suggests that traditional methods of instruction produce greater success in domain specific areas [8] than those that assist students in developing techniques that they can use in multiple-topic areas.

This paper describes a response-driven Web-based assessment system (RDWBAS) enhancing learning. The system contains tests assessing recall of facts, high-level thinking, asking students to evaluate consequences and draw conclusions, and various quizzes. RDWBAS responds to students' needs as they progress through the system. This is incorporated by providing a structured learning system environment, suitable for offering automatic help through logic-based and qualitative reasoning mechanisms and guiding students in their navigation. Previous research indicates that interactive-engagement classes can

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 749–755, 2005.

achieve at higher levels than more didactic classes, and that students in active-engagement computer-based physics classes outperform students who receive traditional instruction [12].

A RDWBAS assessing a student's knowledge asks her to attach the level of confidence she has in her answers and then gives recommendations on how to progress through the system. Assessment is most effective when it is ongoing and reflects an understanding of learning as multidimensional, integrated, and revealed in performance over time. RDWBAS enables a lecturer to assess, plan, deliver and manage students' learning by delivering the right learning to the right audience in the right method and thus achieve the right results.

2 Goals

Our main objectives are to provide a simple but effective way of assessing learners' progress by

- tracking every student's progress
- increasing students' perceptions of control
- providing individual instructions for every student
- improving students' problem solving abilities and build higher-level skills
- fostering excellence in the development of new knowledge
- using the results of the assessments to improve student learning
- motivating students

3 Related Work

Research-based good practice addressing the pedagogical, operational, technological, and strategic issues faced by those adopting computer-assisted assessment is described in [11] and [10]. Integrating assessment and instruction is discussed in [7].

A level-based instruction model is proposed in [14]. A model for student knowledge diagnosis through adaptive testing is presented in [5]. An approach for integrating intelligent agents, user models, and automatic content categorization in a virtual environment is presented in [15]. A computer system generating interactive dialogs in a mathematics teaching application is described in [13]. A model for detecting student misuse of ITS is presented in [1].

Permutational multiple choice question tests have been used for assessing high-level thinking [3]. Usually the student is asked to consider two similar concepts, or two complementary taxonomies. A question is answered correctly if each stem (questions or incomplete sentences) is matched up with the appropriate key (correct option). This implies only one correct answer and no differentiation is made between a wrong answer caused by miscalculation, a wrong answer caused by lack of conceptual thinking, or application of a wrong method.

A grading system that does not make a distinction between a question followed by a wrong answer and a question that is not answered implies that a student is 100% confident in her answer. Being able to properly judge the confidence of one's answers is an important part of being knowledgeable [4].

4 A Learner-Centered Approach

Student knowledge is critical for individualizing the instructional process. RDWBAS provides different tests with respect to levels of difficulties. At any level RDWBAS presents students with positive examples reinforcing understanding and negative examples establishing conceptual boundaries.

Each level has tutorial material generated for it; since it is important to target tutorial tasks at the student's ability, this is seen as being of more educational benefit than offering the same tutorials to all students and then assigning a student to a level based upon the grade the student achieves [9]. It is incorporated by including different help functions. Intelligent agents provide different students with different pages according to their needs. Additional explanations and examples helping to clear current difficulties and misconceptions are provided by RDWBAS without human tutors.

4.1 Assessing High-Level Thinking

Students' conceptual thinking can be assessed by presenting them with tests where all the correct answers should be chosen and/or answers require integration of several components or approaches [2] and [3].

A test where the set of putative answers may contain several correct answers and several wrong answers allows for very detailed feedback since it provides increased accuracy, reliability, and usability. Each one of the unique answer combinations to every question implies a different level of knowledge about the topic. There is no restriction on the number of correct and wrong alternatives, which prevents students from 'intelligent' guessing about the number of correct answers following each question.

Tests are designed to assess critical thinking applying Bloom's Taxonomy. Such tests contain stems asking students to identify the correct outcome of a given circumstance, map the relationship between two items into a different context, respond to what is missing or needs to be changed within a provided scenario, and evaluate the proposed solution based upon criteria provided.

4.2 Web Interface

User navigation in an intelligent tutoring system should consider both preventing the user from becoming overwhelmed with information and losing track of where she is going, while still permitting her to make the most of the facilities the system offers.

To decide when to recommend a student to solve another problem, it is necessary to assess the student's knowledge. This is achieved in an automatic way allowing students to work without a human tutor. This is realized by including a number of tests, whose role is to judge the student's level of knowledge and understanding of a subject.

A subject is a collection of elements $(E_1, E_2, ..., E_n)$. The order of the elements indicates that understanding of an element with a higher number requires understanding of the elements with lower numbers. Suppose a student is working with an element E_i (Fig. 1). At a place chosen by the instructor, she receives a request to take a multiple choice test. The student is asked to mark all the correct answers to every question and to attach a level of confidence (20%, 40%, 60%, 80%, 100%). The system then provides a diagnostic report detecting miscalculations, misconceptions, and lack of knowledge. Based on the recommendations of this report the student is advised to proceed with the next element, take another test on the same element, or work with a selection of previous elements. The system guides the student downwards according to her/his mistakes and brings her/him upwards using the same path without a human tutor. However, students have an opportunity, based on their own judgment, to place themselves in any of the elements. By tracking the student's path automatically in response to the student's answers, the system helps the student to learn from her/his mistakes. The system provides information about the level of knowledge of each

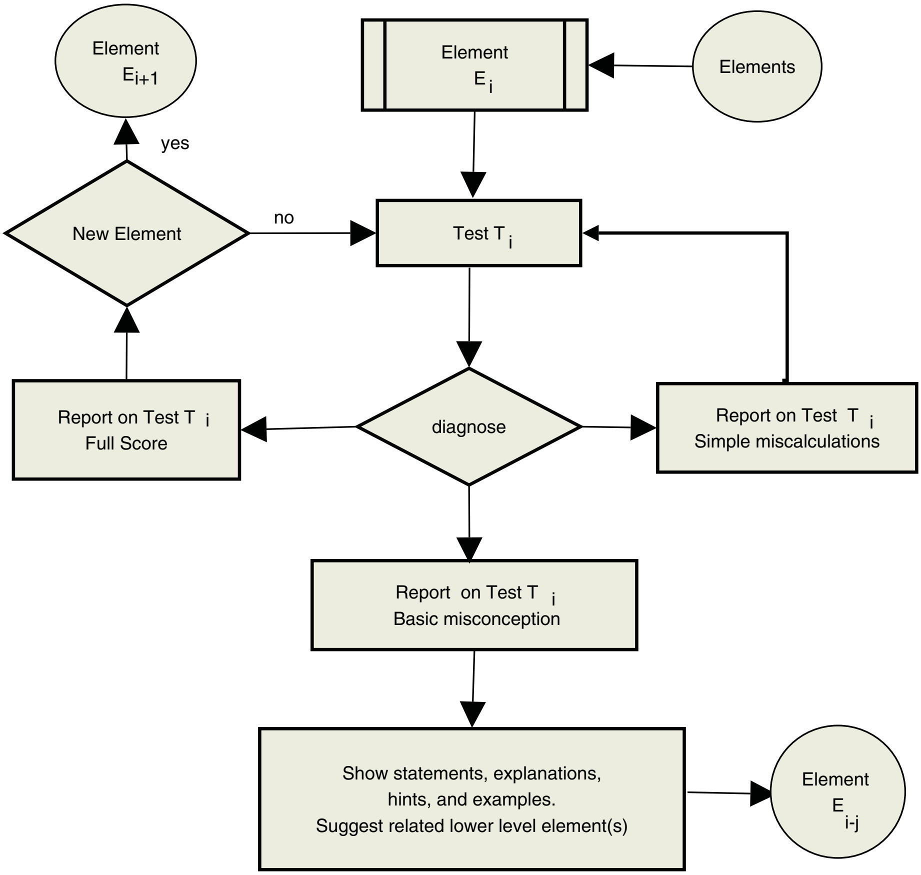


Fig. 1. The system

registered user on every particular element, as well as information on the amount of students going back to a particular problem or/and statement. This helps content developers to improve certain parts by including further explanations and better examples.

5 System Architecture

The system framework is composed of three main software components.
Web server: Apache Web server is used to provide users with dynamic contents, test Web forms and the user's navigational structure.
DBMS: PostgreSQL relational database is used to store the assessment system structural definition in XML data, elements metadata, tests metadata, user administration, user status, and user profile.
Script program modules: Python is used to provide a server-side scripting environment for dynamic contents, database integration, intelligent diagnostics and integration between system components.

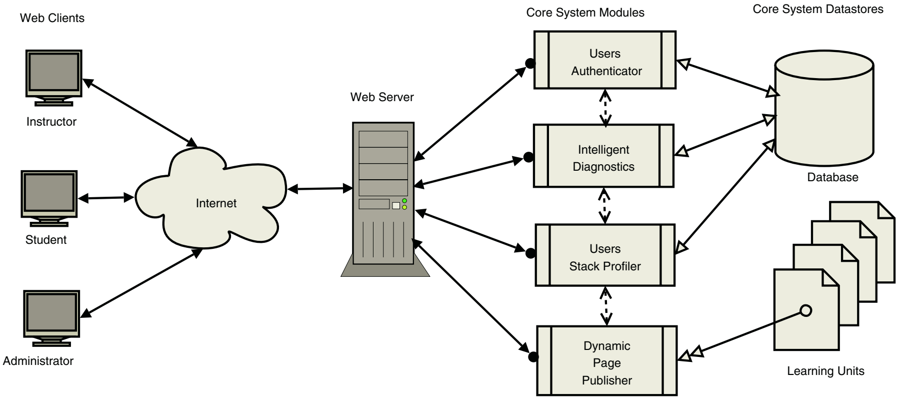


Fig. 2. System architecture

The assessment system structure is defined by pedagogical requirements. This structure defines dependencies among elements, levels and relationships between tests options, and inference rules used in the diagnostic sub-systems. This structure is crucial in providing the learner with a personalized learning work-flow for efficient learning. Each element is a self-contained learning unit with a set of tests assigned to it. The user's responses to Web-based tests provide the system with necessary data. The intelligent diagnostic sub-system analyzes these data using the programmed inference rules and provides the learner with an immediate recommendation on how to proceed. The test score and user status are saved in the database.

In the recommendation on how to proceed, a user can choose to subscribe to one or more suggested elements. The user's element subscribtions are placed in a stack-like structure in the user profile data. The system presents the user with the top most element in each new learning session.

Initially, the profile stack contains a sequential ordering of elements in a given subject. A user can choose to skip any presented element and go to the next one at any time. A user is considered to have completed a course when her/his profile stack is empty and she/he has passed all compulsory tests assigned to the course.

All element names taken by the user during the course and scores of the tests are saved in the user audit-trail. Such audit-trail data is used for billing purposes while global analysis of the course and feedback data is used to improve contents and tests for each subject.

6 Experience Using RDWBAS

Course grades for two groups of undergraduate engineering students in calculus have been compared. The control group had no access to Web-based assessments. The experimental group had access to Web-based assessments. The control group contained 96 students enrolled in the course during the Fall semester 2003. The experimental group contained 93 students enrolled in the course during the Fall semester 2004. The results for the experimental group were consistent with their performance in the course, where those who scored high overall results on the final exam also scored high marks for their Web-based tests.

7 Conclusion and Future Work

A central problem of learning at any moment is to determine what the student knows and does not know, and to offer appropriate help based on that knowledge. The presented RDWBAS responds to students' needs as they progress through the system. A structured learning system environment capable of altering some part of the instructional process on an individual basis by the use of individual student profiles is provided.

In deciding when to recommend another question for a student, it is necessary to assess the student's knowledge and the level of confidence she has in her answers. If a student fails a test, she is provided with a diagnostic report showing her weaknesses. This is achieved in automatically allowing students to work without a human tutor.

The results obtained guarantee the usefulness of a full-scale implementation of a Web-based assessment model in the future. Both lecturers and students expressed a desire to expand their usage of computer technology.

The formal evaluation elicited useful perceptions concerning the effectiveness of the way in which Web-based assessments were integrated into lectures and tutorials.

As a future work, recommendations arising from this formal evaluation will be used for redesigning features in the next version of these assessments. In our next phase of experiments we will evaluate the suitability of the rules already implemented and the effectivity over time of the system.

References

1. Baker, R.S., Corbett, A.T., Koedinger, K.R.: Detecting student misuse of intelligent tutoring systems. Lecture Notes in Computer Science, Vol. 3220. Springer-Verlag, Berlin Heidelberg New Jork (2004) 531–540
2. Bush, M. : A multiple choice test that rewards partial knowledge. Journal of Further and Higher Education **25**2 (2001) 157–163
3. Farthing, D.W.: Permutational multiple choice questions: An objective and efficient alternative to essay-type examination questions. Proc. of 3rd Annual Conference on Integrating Technology into Computer Science Education (ITiCSE'98) (1998)
4. Gardner-Medwin, A.R.: Confidence assessment in the teaching of basic science. Association for Learning Technology Journal **3** (1995) 80–85
5. Guzmàn, E., Conejo, R.: A model for student knowledge diagnosis through adaptive testing. Lecture Notes in Computer Science, Vol. 3220. Springer-Verlag, Berlin Heidelberg New Jork (2004) 12–21
6. Janvier, W.A., Ghaoui, C: Using Communication Preference and mapping Learning Styles to Teaching Styles in the Distance Learning Intelligent Tutoring System-WISDeM. Lecture Notes in Artificial Intelligence, Vol. 3190. Springer-Verlag, Berlin Heidelberg New Jork (2003) 185–192
7. Jensen, M.R., Feuerstein, R.: The learning potential assessment device: From philosophy to practice. In C.S. Lidz (Ed.), Dynamic assessment: An interactional approach to evaluating learning potential. New York, Guilford Publications, Inc. (1987) 379–402
8. de Jong, T., Ferguson-Hessler, M.G.M.: Types and qualities of knowledge. Educational Psychologist **31** (1996) 105–113
9. Hartley, J.R.: Interacting with multimedia. University computing **15** (1993) 129–136
10. Harper, R.: Correcting computer-based assessments for guessing. Journal of Computer Assisted Learning **19** (2003) 2–8
11. Hirsh, L., Saeedi, M., Cornillon, J., Litosseliti, L.: A structured dialogue tool for argumentative learning. Journal of Computer Assisted Learning **20**(1) (2004) 72–80
12. Huffman, D, Goldberg, F., Michlin, M.: Using computers to create constructivist environments: impact on pedagogy and achievement. Journal of Computers in mathematics and science teaching **22**(2) (2003) 151–168
13. Mora, M.A., Moriyòn, R., Saiz, F.: Role-based specification of the behavior of an agent for the interactive resolution of mathematical problems. Lecture Notes in Computer Science, Vol. 3220. Springer-Verlag, Berlin Heidelberg New Jork (2004) 187–196
14. Park, C., Kim, M.: Development of a Level-Based Instruction Model in Web-Based Education. Lecture Notes in Artificial Intelligence, Vol. 3190. Springer-Verlag, Berlin Heidelberg New Jork (2003) 215-221
15. Santos, C.T., Osòrio, F.S.: Integrating intelligent agents, user models, and automatic content categorization in virtual environment. Lecture Notes in Computer Science, Vol. 3220. Springer-Verlag, Berlin Heidelberg New Jork (2004) 128–139

Intelligence-Based Educational Package on Fluid Mechanics

KwokWing Chau

Department of Civil and Structural Engineering, Hong Kong Polytechnic University,
Hunghom, Kowloon, Hong Kong
cekwchau@polyu.edu.hk

Abstract. From the student feedback questionnaire, some students opined that the concepts of fluid mechanics are quite abstract and that they have difficulty in grasping the phenomena in real life situation. Hence, it demands some innovative learning methodologies to help arouse their interest. This paper depicts the development and implementation of an interactive teaching package on learning of basic fluid mechanics with a knowledge-based system approach. The prototype package is designed to guide engineering students in self-directed learning through the processes of interaction, reflection, and application, thus furnishing an opportunity of stimulating pedagogical environment. Diagnostic assessment is undertaken for every scenario of possible prompted answer on a specific topic, so as to evaluate the most probable shortfall or misconception of that particular student.

1 Introduction

From the student feedback evaluation, some students in civil engineering and mechanical engineering fields opined that the concepts of fluid mechanics, which involves the fundamental principles of physical science and applied mathematics, are too abstract and that they have difficulty in fully grasping the real phenomena. Hence, some innovative teaching and learning methodologies are necessary to help arouse their interest. With the advancements in the fields of computer and education technology, innovative learning package is becoming a general trend [1-2]. It appears that new technology will build a new paradigm on education, with self-directed learning as a foundation strategy. It has the potential for effecting fundamental changes in the design of pedagogical processes and the instructional system. Moreover, the traditional learning and training system is sometimes considered not effective enough since it offers little facility in tracking the progress of the student. It is suggested that simply to present materials to the students is not enough.

Recent advancements in artificial intelligence technology have made it possible for computer programs, by encoding knowledge and reasoning, to simulate human expertise in narrowly defined domains during the problem-solving process. There exist the most important trends in intelligence-based educational systems (IBES), which comprise various intelligent technologies such as knowledge-based system (KBS), fuzzy systems, neural networks, genetic algorithms, artificial immune systems, and their implementation in a multi-agent framework and in form of hybrid intelligent systems. These smart interactive tutoring systems are actually required and are able to fit themselves to the student's individual unique needs [3-4]. Being one form of IBES, a KBS

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 756–762, 2005.

is capable to incorporate systematically the heuristic knowledge and expertise. By knowledge processing facilities, individual expert's knowledge could be stored under rule frame on a permanent basis so long as such rules are valid and update of such knowledge base whenever necessary is accomplishable over passage of time. The progress and development of KBS suggests that "machine expert" can play a vital role in decision making. It has been proven to be appropriate in furnishing solutions to domain problems that require considerable rules of thumb, judgment or expertise, in particular under the following types of classification, namely, education, diagnosis, interpretation, planning, and design. KBS has made widespread applications in a variety of domain problems and is proven to be capable of attaining a standard of performance comparable to that of a human expert [5-25]. It is towards this direction that the present study goes for extensive knowledge base for teaching and learning of fluid mechanics. In this paper, the development and implementation of a prototype interactive teaching package on learning of basic fluid mechanics with a KBS approach is delineated. Figure 1 shows the block diagram of this intelligence-based tutoring system.

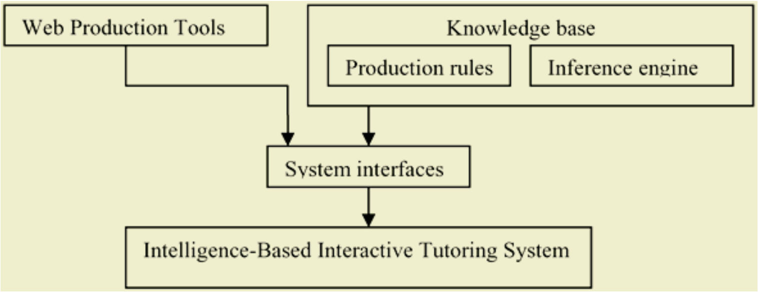


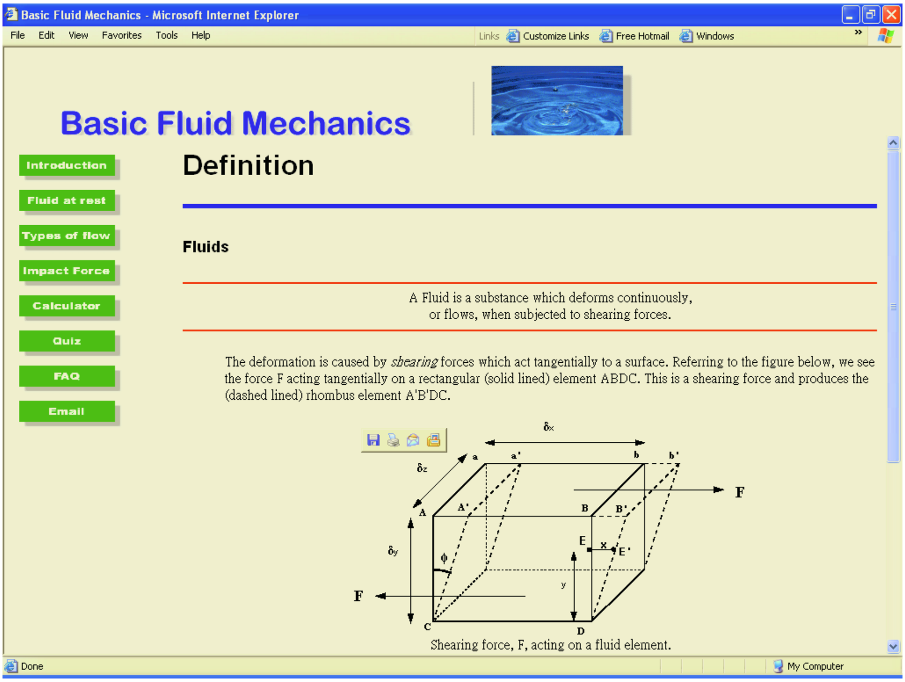
Fig. 1. Block diagram of the intelligence-based tutoring system

2 Use of Technology on Education

Nowadays, an expanding multimedia communication system offers the advantage of furnishing diversified and enhanced delivery mechanisms of quality education. However, effective instruction with technology must be driven by sound pedagogical principles, involving critical thinking, and providing a real community to students. With the increasing quality and availability of technology, learning has become rapid, effective, flexible, and convenient. In addition, it furnishes the immediacy and range of interaction comparable with face-to-face learning. It is apparent that the groups best served by technology are individuals who have special demands, have family responsibilities, and work and reside in remote areas. Moreover, there is stronger demand on educational institutions to be more efficient, to improve in areas including instructional quality, mode of study, access, and costs. There exists strong demand for higher education to become more convenient, flexible, and effective for these individuals.

Moreover, it permits the students to acquire quality learning experience to suit their specific demands or capabilities. Students can freely and directly gain access to various parts of the course contents, and if they envisage any queries at any stage, they can point straight back into the relevant sections or into the references and back again.

As such, it furnishes a dynamic and active learning environment and provides an opportunity of stimulating pedagogical environment to take care of engineering students in self-directed learning through interaction, application and reflection.



Basic Fluid Mechanics

Definition

Fluids

A Fluid is a substance which deforms continuously, or flows, when subjected to shearing forces.

The deformation is caused by *shearing* forces which act tangentially to a surface. Referring to the figure below, we see the force F acting tangentially on a rectangular (solid lined) element ABDC. This is a shearing force and produces the (dashed lined) rhombus element A'B'DC.

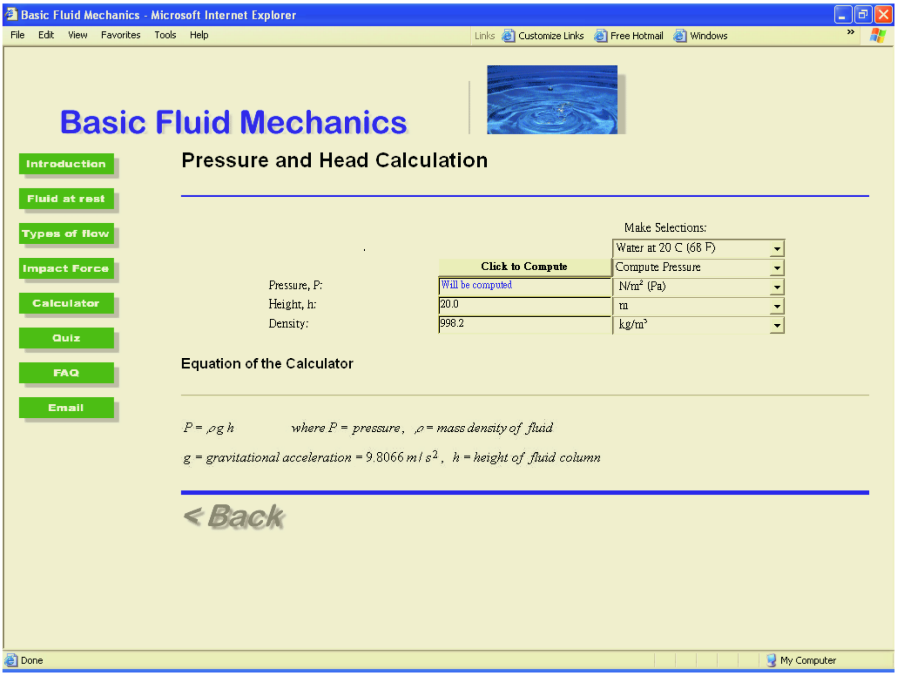
Shearing force, F, acting on a fluid element.

Fig. 2. Screen displaying instructional knowledge on basic fluid mechanics

Queries have been raised regarding the pedagogical quality that technology furnishes. Some educators may be concerned that teaching package is neither personal nor interactive and is consequently less effective than face-to-face instruction. A serious criticism is that it fails to create an effective learning environment due often to poor design. The common weakness of many learning packages is their misapprehension that information is equal to learning and material is laid out basically on the package in a regular textbook format. In such cases, learners are merely passively involved in electronic page turning when reading and sorting through material.

Novel technologies may provide flexible as well as prosperous media for representing what students know and what they are learning. Yet they should function as intellectual tool kits that assist learners in establishing meaningful personal interpretations and representations of their environment. Hence, this new learning package is founded on learner demands for quality content, delivery, and service that leads to desired learning outcomes. It comprises a more constructivist view where learners are encouraged to reorganize, manipulate and personally synthesize course materials under an active and dynamic environment.

It necessitates a shift in teaching paradigm where instructors may need to adopt a more student-centered approach to their teaching. The emphasis is now placed on collaboration and active learning. The contexts should cater for learners of wide-ranging perspectives. Interactions are significant in that they render participation in the cycle of instruction, training, performance assessment, and improvement processes. They enable learners to tailor learning experiences to meet their specific capabilities or demands. Interactions allow clarification and the transfer of new ideas to extant conceptual frameworks.


Basic Fluid Mechanics - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links Customize Links Free Hotmail Windows
Introduction
Fluid at rest
Types of flow
Impact Force
Calculator
Quiz
FAQ
Email
< Back
Done
My Computer

Basic Fluid Mechanics

Pressure and Head Calculation

Make Selections:

Water at 20 C (68 F)

Click to Compute | Compute Pressure

Pressure, P: Will be computed | N/m² (Pa)

Height, h: 20.0 | m

Density: 998.2 | kg/m³

Equation of the Calculator

$P = \rho g h$ where P = pressure, ρ = mass density of fluid

g = gravitational acceleration = $9.8066\ m/s^2$, h = height of fluid column

Fig. 3. Screen displaying the interactive "What-if" analysis on pressure and head

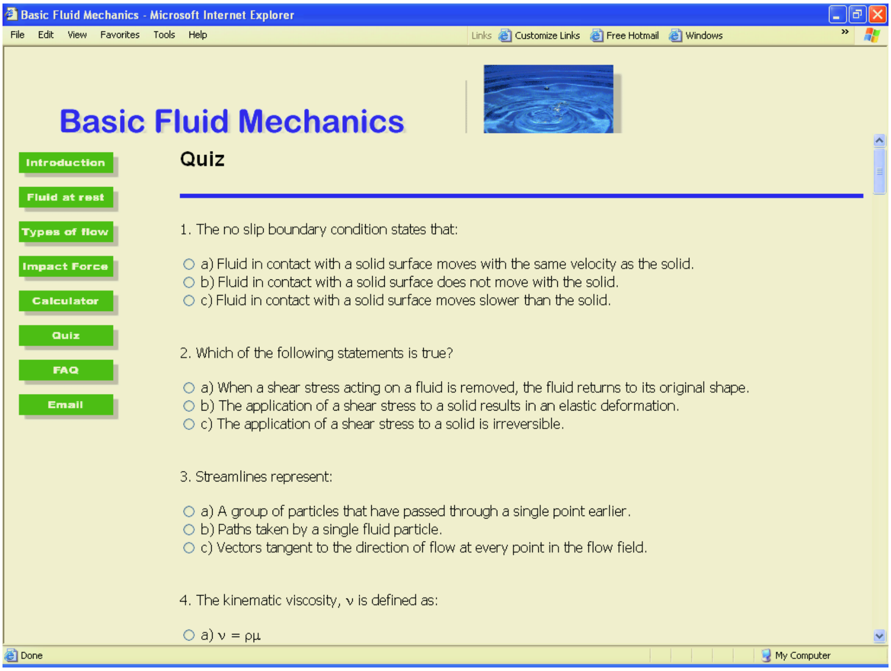
Table 1. Evaluation of the package on the basis of student feedback questionnaire survey

Questionnaire item	Average rating#
The presented material is relevant to the domain subject.	4.2
The package is easy to comprehend and greatly accessible.	4.3
The package is very helpful in understanding the topic.	4.1
The system is interactive and user-friendly.	4.3
The material with multiple formats of presentation is interesting.	4.1
The tool greatly arouses their interest in this subject.	4.2
Users can actively control the pedagogical process via the tool.	4.0
Users are proficient in using computer.	3.5

#5 = Strongly Agree, 4 = Agree, 3 = Neutral, 2 = Disagree, 1 = Strongly Disagree

3 Development Shell

This system has been developed and implemented using a microcomputer-based KBS shell Visual Rule Studio in order to facilitate development of the knowledge base on fluid mechanics. It is a hybrid application development tool under object-oriented programming design environment. This shell acts as an ActiveX Designer under the Microsoft Visual Basic 6.0 programming environment. Both production rules and procedural methods are employed to represent standard and heuristic knowledge on fluid mechanics. Rules are isolated as component objects, which are separated from both objects and application logic. As such, it produces objects that can interact with most modern development software.



Quiz

1. The no slip boundary condition states that:

○ a) Fluid in contact with a solid surface moves with the same velocity as the solid.
○ b) Fluid in contact with a solid surface does not move with the solid.
○ c) Fluid in contact with a solid surface moves slower than the solid.

2. Which of the following statements is true?

○ a) When a shear stress acting on a fluid is removed, the fluid returns to its original shape.
○ b) The application of a shear stress to a solid results in an elastic deformation.
○ c) The application of a shear stress to a solid is irreversible.

3. Streamlines represent:

○ a) A group of particles that have passed through a single point earlier.
○ b) Paths taken by a single fluid particle.
○ c) Vectors tangent to the direction of flow at every point in the flow field.

4. The kinematic viscosity, ν is defined as:

○ a) $\nu = \rho\mu$

Fig. 4. Screen displaying quiz on basic fluid mechanics

4 The Prototype Teaching Package

The emphasis of this study is on the diagnostic assessment of learning performance and on the ensuing learning directive designed by the intelligent system, which depends on the response of the learner and the assessment outcome. Assessment exercises are carefully designed for each selected topic in fluid mechanics, covering all possible answers from the learner in mind. The covered topics include fluid at rest, types of flow, impact force, similitude, pipe flow, open channel flow, hydrology, hydrodynamics, coastal hydraulics, unsteady flow, and wind loading on structures. For each scenario of prompted answer from the student, diagnostic assessment is performed by the system to determine the most probable shortfall or misconception of the specific student on that particular topic. This heuristic knowledge can be represented by knowledge rules under the KBS approach. Figure 2 shows the screen displaying instructional knowledge on basic fluid mechanics. Figure 3 shows the screen displaying the interactive "What-if" analysis on pressure and head. Figure 4 shows screen displaying quizzes on basic fluid mechanics.

The prototype package is evaluated on the basis of a student feedback questionnaire survey. Table 1 shows some preliminary findings of the users' feedback on several useful points regarding the scope and effectiveness of this package. From the results, it can be concluded that the tool is found to be very relevant, easy to comprehend, accessible, helpful in understanding, interactive, user-friendly, interesting, arousing their interest and highly controllable although their computing literacy is not too high.

5 Conclusions

In this paper, the development and implementation of a prototype interactive teaching package on learning of basic fluid mechanics with a KBS approach is depicted. It is demonstrated that various theories on hydrology, fluid motion, etc. can be performed using this package through an active and dynamic learning environment. The flexibility and open infrastructure of the package have been shown to be able to act as a media for developing learning application. It offers the possibility of providing a stimulating learning environment to engage learners in meaningful learning through reflection, application, and interaction. The engineering students can gain deeper insight on this abstract subject through the interaction furnished in this package.

References

1. Berge, Z.L.: Guiding Principles in Web-Based Instructional Design. Educational Media International **35(2)** (1998) 72-76
2. Wiens, G., Gunter, G.A.: Delivering Effective Instruction via the Web. Educational Media International **35(2)** (1998) 95-99
3. Negoita, M.G., Pritchard, D.: Developing a "Virtual Student" Model to test the Tutor and Optimizer Agents in an ITS. Lecture Notes in Computer Science **3213** (2004) 240-252
4. Negoita, M.G., Pritchard, D.: An Optimizer Agent that empowers an ITS system to "On-the-Fly" modify its Teaching Strategies. Lecture Notes in Computer Science **3213** (2004) 914-921
5. Albermani, F., Chau, K.W.: Web-Based Knowledge-Based System on Liquid Retaining Structure Design as Instructional Tool. Lecture Notes in Computer Science **2436** (2002) 95-105
6. Chau, K.W.: An Expert System for the Design of Gravity-type Vertical Seawalls. Engineering Applications of Artificial Intelligence **5(4)** (1992) 363-367
7. Chau, K.W.: A Prototype Knowledge-Based System on Unsteady Open Channel Flow in Water Resources Management. Water International **29(1)** (2004) 54-60
8. Chau, K.W.: Knowledge-Based System on Water-Resources Management in Coastal Waters. Water and Environment Journal **18(1)** (2004) 25-28
9. Chau, K.W., Albermani, F.: Expert System Application on Preliminary Design of Liquid Retaining Structures. Expert Systems with Applications **22(2)** (2002) 169-178
10. Chau, K.W., Albermani, F.: A Coupled Knowledge-Based Expert System for Design of Liquid Retaining Structures. Automation in Construction **12(5)** (2003) 589-602
11. Chau, K.W., Albermani, F.: Knowledge-Based System on Optimum Design of Liquid Retaining Structures with Genetic Algorithms. Journal of Structural Engineering ASCE **129(10)** (2003) 1312-1321
12. Chau, K.W., Albermani, F.: Hybrid Knowledge Representation in a Blackboard KBS for Liquid Retaining Structure Design. Engineering Applications of Artificial Intelligence **17(1)** (2004) 11-18

13. Chau, K.W., Albermani, F.: An Expert System on Design of Liquid-Retaining Structures with Blackboard Architecture. Expert Systems **21(4)** (2004) 183-191
14. Chau, K.W., Albermani, F.: A Knowledge-Based System for Liquid Retaining Structure Design with Blackboard Architecture. Building and Environment **40(1)** (2005) 73-81
15. Chau, K.W., Anson, M.: A Knowledge-Based System for Construction Site Level Facilities Layout. Lecture Notes in Artificial Intelligence **2358** (2002) 393-402
16. Chau, K.W., Chen, W.: An Example of Expert System on Numerical Modelling System in Coastal Processes. Advances in Engineering Software **32(9)** (2001) 695-703
17. Chau, K.W., Cheng, C., Li, C.W.: Knowledge Management System on Flow and Water Quality Modeling. Expert Systems with Applications **22(4)** (2002) 321-330
18. Chau, K.W., Cheng, C.T., Li, Y.S., Li, C.W., Wai, O.: An Intelligent Knowledge Processing System on Hydrodynamics and Water Quality Modeling. Lecture Notes in Artificial Intelligence **2358** (2002) 670-679
19. Chau, K.W., Cheung, C.S.: Knowledge Representation on Design of Storm Drainage System. Lecture Notes in Computer Science **3029** (2004) 886-894
20. Chau, K.W., Ng, V.: A Knowledge-Based Expert System for Design of Thrust Blocks for Water Pipelines in Hong Kong. Journal of Water Supply Research and Technology - Aqua **45(2)** (1996) 96-99
21. Chau, K.W., Sze Y.H.: AI-Based Teaching Package for Open Channel Flow on Internet," Lecture Notes in Computer Science **3143** (2004) 98-104
22. Chau, K.W., Yang, W.W.: A Knowledge-Based Expert System for Unsteady Open Channel Flow. Engineering Applications of Artificial Intelligence **5(5)** (1992) 425-430
23. Chau, K.W., Yang, W.W.: Development of an Integrated Expert System for Fluvial Hydrodynamics. Advances in Engineering Software **17(3)** (1993) 165-172
24. Chau, K.W., Yang, W.W.: Structuring and Evaluation of VP-Expert Based Knowledge Bases. Engineering Applications of Artificial Intelligence **7(4)** (1994) 447-454
25. Chau, K.W., Zhang, X.Z.: An Expert System for Flow Routing in a River Network. Advances in Engineering Software **22(3)** (1995) 139-146

Generalized Composite Motif Discovery

Geir Kjetil Sandve and Finn Drabløs

Norwegian University of Science and Technology, 7052 Trondheim, Norway
{Sandve,Finn.Drablos}@ntnu.no

Abstract. This paper discusses a general algorithm for the discovery of motif combinations. From a large number of input motifs, discovered by any single motif discovery tool, our algorithm discovers sets of motifs that occur together in sequences from a positive data set. Generality is achieved by working on occurrence sets of the motifs. The output of the algorithm is a Pareto front of composite motifs with respect to both support and significance. We have used our method to discover composite motifs for the AlkB family of homologues. Some of the returned motifs confirm previously known conserved patterns, while other sets of strongly conserved patterns may characterize subfamilies of AlkB.

1 Introduction

Motif discovery in DNA and protein sequences is an important field in bioinformatics. Unique motifs found in a set of related sequences are often associated with the biological activity of the sequences. Motifs representing active site residues in enzymes (proteins) or transcription factor binding sites in genomes (DNA) are typical examples. Such motifs can also be used for classification of novel sequences or sequences outside the original training set. Both probabilistic and deterministic approaches are used. Arguably, deterministic approaches give the most easily interpretable results, as they represent motifs e.g. by subsets of regular expressions that either match a given sequence or not.

There are many different algorithms for motif discovery, including manual approaches. The earliest algorithms had very limited expressibility and could only discover substrings of amino acid symbols. PROSITE[1], a database of manually annotated motifs, in many ways set the standard for expressibility of deterministic motifs for proteins. In addition to exact symbols, PROSITE patterns also consist of fixed gaps, flexible gaps and ambigous symbols. Most automated motif discovery tools are only able to discover motifs consisting of a subset of these components.

The discovery of motif combinations is an area of active research, for which both probabilistic and combinatorial approaches are used. Gibbs sampler[2] and PRINTS[3] are two well-known probabilistic approaches. Most combinatorial approaches discover spaced dyads[4][5] or ordered sets of motifs with strong distance constraints[6]. Brazma et al.[7] are among the few methods that discover unordered sets of motifs.

A set of single motifs is a general starting point for composite motif discovery. Many advanced methods exist for the discovery of single motifs, and none are

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 763–769, 2005.

superior in all respects[8]. We have therefore chosen to develop an algorithm for the discovery of motif combinations that can use single motifs generated by any deterministic motif discovery tool.

GCMD (Generalized Composite Motif Discovery)[1], exhaustively identifies the most significant combinations of a set of precomputed motifs. It can be set to discover both ordered and unordered motifs, with or without distance constraints. In addition to being flexible with regards to both single and composite motif model, and exhaustive in search for combinations, two properties clearly distinguish our algorithm from previous approaches: we model the problem as a two-goal optimization with the optimal Pareto front as output, and we automatically discover potential subfamilies.

GCMD is here discussed mainly in terms of protein sequence motifs. However, the tool itself is general and can also be applied to motifs from DNA sequences.

2 The Generalized Composite Motif Discovery Tool

In broad terms, GCMD takes as input a set of single motifs and exhaustively discovers the optimal motif combinations with respect to both support and significance. This is more thoroughly explained in the following sections.

2.1 Vocabulary

The set of sequences that have at least one occurrence of a given motif, is called the *occurrence set* of the motif. The cardinality of the occurrence set is referred to as *support*.

We use the term *single motifs* to denote the motifs that are input to the GCMD algorithm, and *composite motifs* to denote the discovered motifs that are sets of single motifs. The term *component* is used to denote one of the single motifs that makes up a composite motif.

We also use the terms *Pareto domination* and *Pareto front* in multiple criteria optimization. A motif is Pareto dominated if there exists another motif having equal or higher values of both support and significance, where one of the values is strictly higher. Since support is a discrete value, this means that a motif is Pareto dominated if there exist another motif with equal or higher support, and strictly higher significance. The Pareto front is the set of all non-dominated motifs. In our case this is the most significant motif for each value of support.

2.2 Motif Representations

The first step in using GCMD is to discover deterministic single motifs with a separate motif discovery tool. For tools that discover probabilistic motifs, a threshold may be used to make them deterministic. A bitstring is then constructed for each motif, where the i'th bit is 1 if the motif has an occurrence in

[1] The code is available upon request to first author

the i'th sequence, and 0 otherwise[7]. A composite motif occurs in a sequence if, and only if, every single motif in the set occurs in the sequence. This leads to a basic representation of a composite motif as a set of indexes to its component motifs, as well as an occurrence set calculated by taking the intersection of the occurrence sets of all component motifs.

2.3 Significance Evaluation

Significance of motifs is measured as negative log-likelihoods, using the same calculations as the motif discovery method Splash[9]. More specifically, the significance of a single motif is the negative log-likelihood of observing the motif in a random background sequence with the same amino acid distribution as the input sequences. As single motifs usually are short compared to sequence length, the log-likelihood of a composite motif is in general well approximated as the sum of log-likelihoods of its components.

2.4 Significance vs Support

Both significance as well as support is important when evaluating motifs, and it is not easy to make the right trade-off between these properties when doing automated motif discovery. Most algorithms require a threshold on support, and this threshold is often user specified. Using a very strict value may lead to loss of significant motifs that are characteristic of subfamilies of sequences. On the other hand, a too permissive threshold may lead to searches dominated by motifs with high statistical significance in subsets of sequences, and one may lose less significant motifs representing weak commonalities characteristic of larger sequence families.

By formulating the motif discovery problem as a two goal optimization, we can explore a very large search space of interesting motifs, and return information about this in a condensed form as a Pareto front. The user gets a diverse set of motifs, and can readily see the tradeoff between significance and support as the number of sequences taken into consideration increases. This removes the need to set explisit thresholds on support or significance.

2.5 Pruning of Search Space

GCMD traverses the search space exhaustively and returns the set of Pareto optimal composite motifs. The size of the search space is $\binom{n}{c}$, where n is the number of single motifs used as input to GCMD, and c is the desired number of components in the composite motifs. Many algorithms exist for the mining of frequent item sets. Brazma et al.[7] uses the algorithm of Toivonen[10] to discover unordered sets of motifs. As this algorithm do pruning only based on support, it can not handle the large number of input motifs and low values of support that we are interested in.

We have developed a branch-and-bound algorithm, tailored to our two goal optimization problem, that is very efficient on real biological data. Since our

goal is to find an optimal Pareto front with respect to support and significance, we need to determine upper bounds on both of these values. An upper bound on support is simply the minimum support of the current components of the composite motif. To introduce an upper bound on significance, we ensure that when a composite motif is expanded, the new component has a lower significance value than all other components of the motif. Note that this does not reduce the set of composite motifs we are able to discover, it only excludes all but one of the $n!$ permutations that corresponds to the same combination of n single motifs. For a given motif c_i, this leads to a straightforward upper significance bound on any expansions of c_i with n components :
$s(c_n) \leq s(c_i) + (n - i) * s(c_i(i))$, where c_i is a motif with i components, $c_i(i)$ is the i'th component of motif c_i, and $s(c)$ is the significance of motif c.

With these upper bounds in place we can make a recursive function that takes as parameter a composite motif c that is to be expanded. For each single motif s with significance lower than all current component significances of the motif, we check whether the resulting upper bounds on support and significance are dominated by the current Pareto front. If not, a new composite motif is formed from c, with the single motif s as an added component. The resulting motif is stored in the Pareto front if it has reached the desired number of components, otherwise it is again expanded recursively.

In order to reduce the number of explored composite motifs even further, we explore the expansions of a given composite motif in order of decreasing significance of single motifs. Note that the support of the composite motif before any new expansion is an upper bound on support. As the upper bounds are monotonically decreasing, we can stop exploring new expansions of a composite motif as soon as the upper bounds on support and significance are dominated by the Pareto front.

2.6 Automated Subfamily Discovery

The Pareto front of composite motifs for a family may contain significant motifs with relatively low values of support. It is natural to ask whether such a motif characterize a subfamily of the data set. One may therefore try to discover new motifs in the sequences that are not in the occurrence set of the first composite motif. Since the goal is to find motifs that are common to as many sequences as possible, we have restricted automated subfamily discovery to only two subfamilies and also demand that one of the motifs belong to the Pareto front of the whole family. Significance values of motifs are log-likelihoods, and a two-subfamily-motif occur in a given sequence if either of the one-subfamily-motifs occur in the sequence. Therefore, the significance of a 2-subfamily-motif c is well approximated as: $s(c) = log_2(2^{s(c_a)} + 2^{s(c_b)} + 2^{s(c_a)+s(c_b)})$, where c_a and c_b are the one-subfamily-motifs.

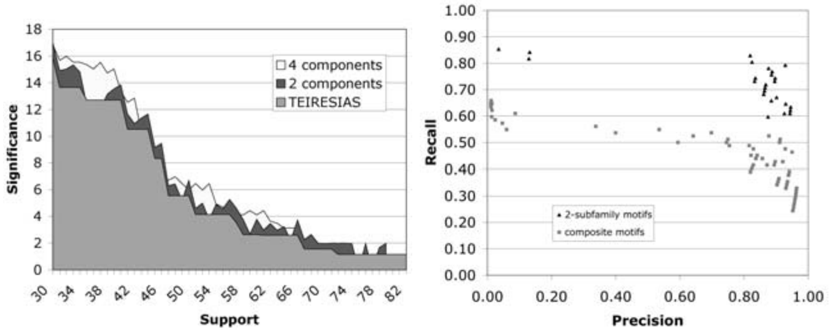
3 Results and Discussion

The family of AlkB homologues (ABHs) was used as a test case for composite motif discovery. The ABHs are members of the 2-oxoglutarate and Fe^{2+}-dependent

(2OG-Fe(II)) oxygenase superfamily[11]. They have been shown to be involved in repair of methylation damage of DNA and RNA through a direct reversal mechanism, where the methyl group is oxidised and spontaneously released as formaldehyde[12]. Recent screening of databases using sensitive search methods has shown that ABH-like sequences are widespread in bacteria and eukaryotes, see Drabløs et al.[13] for a review.

The degree of sequence conservation in the ABH family seems to be very low, basically just a `H.D` motif, an isolated `H` and a `R.....R` motif (using single-letter amino acid symbols) is completely conserved in most ABH alignments. All except the final `R` are involved in coordination of the Fe^{2+} ion, the final `R` is probably involved in substrate binding as it seems to be relatively unique to the ABH family of this superfamily[11]. However, there may be subfamilies within the ABH family with more extensive conservation, and there may be additional conserved patterns in sequence regions that are difficult to align correctly by traditional methods. The ABH family is therefore an interesting test case with practical implications.

A set of 82 AHB-like sequences, previously investigated in [13], was used for the analysis. Teiresias[14] was used to generate 50.000 single motifs from the input sequences, and GCMD was used to identify the Pareto front for composite motifs with 2 and 4 components, using chemical equivalence sets for residue types (Fig. 1(a)). The significance of the composite motifs is higher for most support values compared to single motifs. However, here GCMD is used mainly to identify interesting composite motifs and correlate this with biological significance. The dominating single motif, which is used in most of the composite motifs, is `[ILMV]..H.[DE]`. This corresponds to the first Fe^{2+} binding motif in the ABH sequences. In particular for composite motifs with high support this is mainly combined with variants of the motif `[KR]..[ILMV]..[KR]`, which corresponds to the Fe^{2+} and possibly substrate binding `R` groups. This shows that



(a) Pareto front of single and composite motifs for ABH. Significance is the negative $\log_2$-likelihood of a motif

(b) ROC of the discovered motifs for ABH. Recall is $TP/(TP + FN)$ and precision is $TP/(TP + FP)$)

Fig. 1.

the most interesting motifs identified by GCMD also have biological relevance, and that GCMD is able to identify such motifs from a large and complex set of input data.

However, it is evident that there are subfamilies of ABH-like sequences in the data set, and depending on the selected threshold for support several such subfamilies may be identified. One example is the composite motif `(L..G.[ILMV][ILMV].M....[QN]) & ([FY]....[DE].[ILMV]..H.D)`, which seems to be characteristic of the hABH2/hABH3 subfamily (human ABH type 2 and 3). This subfamily has been extensively studied experimentally[15]. As the detailed 3D structure of the ABH family still has not been experimentally determined, a detailed investigation of the biological relevance of these motifs probably has to be postponed until such data are available. However, this test shows that the GCMD method is able to identify biologically interesting subfamilies in a complex data set.

Although GCMD has not been developed as a classification tool, the classification performance may still serve to validate that the discovered motifs are indeed characteristic for a given family. Fig. 1(b) shows the receiver operating characteristic with respect to recall and precision when using the set of motifs in the Pareto front for classification. The introduction of subfamily motifs leads to a significant improvement in recall, and a larger fraction of the motifs have a high precision, compared to general composite motifs.

The performance of GCMD was also tested on 5 selected families from the PROSITE database. These PROSITE families are assumed to be difficult test cases, as the existing PROSITE patterns give low values for precision and recall. We used TEIRESIAS for single motif discovery. The Pareto front of composite motifs showed an average log-likelihood improvement of 20.4 compared to single motifs. The composite motifs in the Pareto front were used to classify the full set of SWISS-PROT[16] entries. For two of the five families (PS00485, PS00690) we were able to improve both precision and recall as compared to PROSITE, for two families we got comparable performance (PS00732, PS01048), and for the last family the PROSITE motif performed better (PS00187).

4 Conclusion

In our work we have built directly on previous work and focused on finding interesting combinations of single deterministic motifs discovered by separate motif discovery tools. Tests show that our tool is able to identify unique and biologically relevant composite motifs in very large data sets of single motifs.

Future directions of research include expanding the expressibility of deterministic motifs even further, as well as using the tool on other motif discovery problems, like for instance the discovery of transcription factor binding sites.

Acknowledgements

We want to thank Arne Halaas, Kai Trengereid, Magnus Lie Hetland, Rolv Seehuus, Tarjei Hveem and Øystein Lekang for help and suggestions. FD has been supported by FUGE/NFR (151899/150).

References

1. Bucher, P., Bairoch, A.: A generalized profile syntax for biomolecular sequence motifs and its fuction in automatic sequence interpretation. In: Proc Int Conf Intell Syst Mol Biol. 2 (1994) 53–61
2. Neuwald, A. F., Liu, J. S., Lawrence, C. E.: Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Sci. **4** (1995) 1618–1632
3. Attwood, T. K., Beck, M. E., Bleasby, A. J., Parry-Smith, D. J.: PRINTS - a database of protein motif fingerprints. Nucleic Acids Res. **22** (1994) 3590–3596
4. van Helden, J., Rios, A. F., Collado-Vides, J.: Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res. **28** (2000) 1808–1818
5. Eskin, E., Pevzner, P.A.: Finding composite regulatory patterns in DNA sequences. Bioinformatics **18 Suppl 1** (2002) S354–S363
6. Marsan, L., Sagot, M.F.: Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. J Comput Biol. **7** (2000) 345–362
7. Brazma, A., Vilo, J., Ukkonen, E., Valtonen, K.: Data mining for regulatory elements in yeast genome. In: Proc Int Conf Intell Syst Mol Biol. 5 (1997) 65–74
8. Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B.D et al.: Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol **23** (2005) 137–144
9. Hart, R.K., Royyuru, A.K, Stolovitzky, G., Califano, A.: Systematic and fully automated identification of protein sequence patterns. J Comput Biol. **7** (2000) 585–600
10. Toivonen, H.: Discovery of Frequent Patterns in Large Data Collections. PhD thesis, University of Helsinki (1996)
11. Aravind, L., Koonin, E.V.: The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. Genome Biol. **2** (2001) RESEARCH0007
12. Falnes, P.O., Johansen, R.F., Seeberg, E.: AlkB-mediated oxidative demethylation reverses DNA damage in Escherichia coli. Nature **419** (2002) 178–182
13. Drabløs, F., Feyzi, E., Aas, P.A., Vaagboe, C.B., Kavli, B., Bratlie, M.S., Peña-Diaz, J., Otterlei, M., Slupphaug, G., Krokan, H.E.: Alkylation damage in DNA and RNA–repair mechanisms and medical significance. DNA Repair **3** (2004) 1389–1407
14. Rigoutsos, I., Floratos, A.: Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. Bioinformatics **14** (1998) 55–67
15. Aas, P., Otterlei, M., Falnes, P., Vaagboe, C., Skorpen, F., Akbari, M., Sundheim, O., Bjoras, M., Slupphaug, G., Seeberg, E., Krokan, H.: Human and bacterial oxidative demethylases repair alkylation damage in both RNA and DNA. Nature **421** (2003) 859–863
16. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, MC., Estreicher, A. et al.: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31** (2003) 365–370.

Protein Motif Discovery with Linear Genetic Programming

Rolv Seehuus

Norwegian University of Science and Technology
Sem Selandsv 7–9, NO-7034, Norway
rolvinge@idi.ntnu.no

Abstract. There have been published some studies of genetic programming as a way to discover motifs in proteins and other biological data. These studies have been small, and often used domain knowledge to improve search. In this paper we present a genetic programming algorithm, that does not use domain knowledge, with results on 44 different protein families. We demonstrate that our list-based representation, given a fixed amount of processing resources, is able to discover meaningful motifs with good classification performance. Sometimes comparable to or even surpassing that of motifs found in a database of manually created motifs. We also investigate introduction of gaps in our algorithm, and it seems that this give a small increase in classification accuracy and recall, but with reduced precision.

1 Introduction

Motifs, described by different subsets of regular expressions, are important tools in biology used to describe conserved regions in biological sequence data. They can describe important regions with similar function and common ancestral background, and have the virtue of being easy to read and understand. Therefore the automatic discovery of conserved motifs has received quite a bit of attention in the last two decades. Prosite is a database of homologous relationships between proteins, where most of the relationships are described with local motifs [6]. The motifs are manually created and curated by skilled biologists.

The motifs in Prosite contain conserved regions, character classes, and wildcards. Regions of wildcards can have fixed or flexible length as in the following example: `E[XA]...M.{1,3}P{LE}`. The residues in braces represents "one of", and the residues in the curly braces represents "not one of".

In this study, we investigate a Genetic Programming (GP) algorithm using a linear genome (ListGP), and its capabilities to discover interesting motifs in unaligned proteomic data. Including fixed and flexible wildcards in the search increases the resources needed for evaluation of the motifs. Therefore we also investigate how the inclusion of fixed and flexible wildcards affects the results.

To evaluate the population of individuals we employ a piece of specialized hardware called the "Pattern Matching Chip" (PMC) [1]. This yields nearly a *fifty-fold speedup* in our experiments, compared to running regexp matching

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 770–776, 2005.

software like “grep” on an “off the shelf computer” with an Athlon 2700+ CPU and 512Mb of RAM.

In Section 2 we discuss related works and contrast it to our approach. The algorithms we use are described in Section 3. We give results and discussions of experiments in Section 4, and finally some concluding remarks are given in section 5.

2 Previous and Related Work

There has been quite a bit of work performed on motif discovery. Brazma and Jonassen give a good survey of methods based on enumeration and other search methods, and give a view on the different problems in the domain [3]. Evolutionary Algorithms (EA) have been applied to the problem of motif discovery on a few occasions [5, 7]. The experiments have usually been small, presenting results for only a limited number of protein families (usually around two to six.) In addition, they have performed some kind of pre-shaping of the training sets. Either pre-calculated alignments of subsets of the families have been used [8], particularly difficult sequences have been selected for training to simplify search [7], prior knowledge has been used in the shaping of terminal sets [5], or randomly generated data only have been used as negative samples [5, 9].

Koza et al evolved motifs without gaps [7]. In their work the patterns are represented with the standard GP-tree structure, using automatically defined functions (ADFs) for character classes in the motifs.

Hu did some work where he first evolved motif candidates, followed by a local search refinement [5]. He also used a list based genome, where he evolved groups and residues but found the appropriate wildcard regions by a greedy optimization procedure afterwards. Also, as opposed to Koza et al., who evolved the character classes, he stored an index containing classes of residues with similar chemical properties provided by a domain expert. He also used almost the full range of the Prosite language. He reported good results, with motifs very similar to the target Prosite motifs for a few families.

Ross have reported on some experiments for motif discovery in proteins using GP and stochastic regular expressions in both prealigned and unaligned protein sequences [9]. He reports some results for six different protein families. Notably, he has a very small number of positive training samples and he uses the majority of his data points for testing. Like Hu, Ross uses randomly generated sequences as negative training data, instead of sequences or sequence segments from the Swiss-Prot database.

Heddad et al. have very recently used GP to evolve classifiers for the protein targeting problem [4]. They performed experiments using both a grammatical approach to evolve motifs with the full expressibility of the Prosite language, as well as a linear approach to evolve short motifs without gaps but with ambiguous positions. Their motifs were combined to a classifier with different arithmetic expressions. They argued that results might indicate that to solve the protein targeting problem, short conserved motifs of fixed length might be sufficient for building good classifiers.

3 Algorithms and Representations

ListGP is a variant of Genetic Programming [2], a generation based search algorithm. For each generation, a new population of individuals is created through selection, mutation and crossover. We will not elaborate on the mechanics of the algorithm here, as all of it is found in the book by Banzhaf et al. [2]. In this section it therefore suffice to describe the details of our representation, and the crossover and mutation operators performed upon the representation.

In ListGP one individual is represented by a list of nodes, and each node represents either an amino acid residue, a character class (e.g. allowing one position in the motif to match two or more different amino acids) or a wildcard. The initial population of candidates was created by, for each individual, first picking a length sampled from a normal distribution with expected length 8 and standard deviation 4. Afterward each list-element was filled. The length of the motifs found with ListGP was limited to a maximum length of 64 residues (also counting the residues in character classes.)

When creating new nodes (either when initializing or when mutating a position in the list-representation) a single residue is created 60% of the time, whereas character classes are created 40% of the time. When creating a character class, the size of the group is first chosen between two and six. The group is thereafter filled with different random amino acid residues selected uniformly over the alphabet of twenty different amino acids.

To allow for easy handling of contiguous and flexible wild-card regions in our GP-individuals, we introduced two new terminals. Dots,`.`, represented fixed wildcards, and `s` represented an optional wildcard. If `R` represent any group or residue terminal, a motif with flexible gaps is represented with strings like `RRR...s.s..RRR` . This example string can then be translated to a regular expression containing a flexible gap with at least 6 and at most 8 wild-cards by counting the number of different types of contiguous wild-cards there are in an individual: `RRR.{6,8}RRR`

The crossover operator in our experiments is a two-point crossover operator, where the crossover points are selected uniformly over the list in both genomes. Mutations consists of the string edit operations; deletion, replacement or insertion. The operations were selected with width equal probability, and the two latter operations requires the creation of a new node. As for crossover, the position of a mutation is selected uniformly over the genome after an individual is selected for mutation through tournament selection.

All runs of ListGP were done with a population of 1500 individuals for 100 generations. New generations was created by 70% from crossover and 30% by mutation. These parameters were not changed during the experiments. After a run, the motif found with the best score on the training data was evaluated on the test-data.

3.1 Evaluation of Motifs

We used the Matthews correlation coefficient as given in Equation 1 as a fitness measure. In this formula, TP is the number of true positives, TN the true negatives, FP false positives and FN false negatives. C falls in the interval $[-1, 1]$, where a -1 indicates perfect negative correlation and a 1 indicates a perfect positive correlation. If the data have no correlation at all, C equals zero. The Matthews correlation coefficient is a special case of the Pearson correlation coefficient, and a well established measure in biology. It's usage as a fitness measure was inspired from other works [7].

$$C = \frac{TP * TN - FN * FP}{\sqrt{(TN + FN)(TN + FP)(TP + FN)(TP + FP)}} \tag{1}$$

4 Experiments and Results

We selected the protein families to investigate according to two criteria: We wanted the family to contain at least 200 proteins, and the family in Prosite should not be perfect (e.g. it should have more than zero false negatives or false positives.) The first criterion ensured enough positive examples in our test sets, while the second criterion ensured that it was possible to improve on the Prosite motifs. After removing duplicates, 44 different protein families remained. For each family, we labeled all proteins found in the family as positive, and all other proteins in the Swiss-Prot database was labeled as negative.

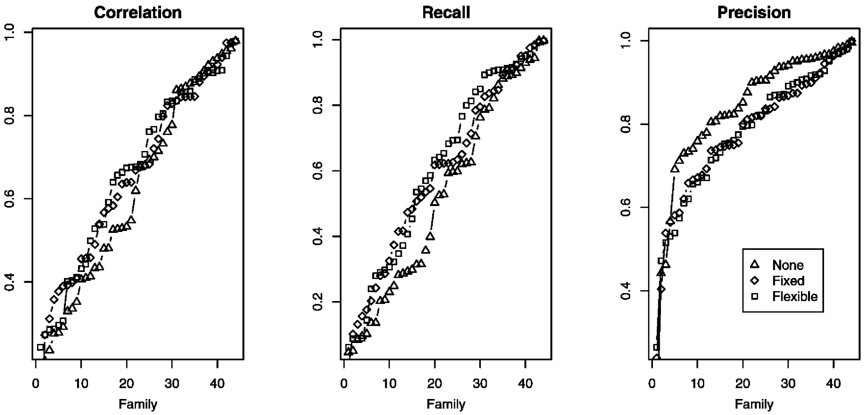


Fig. 1. Comparative view of the different performance measures for the different language subsets

In Figure 1, the evolved motifs are plotted against each other with values sorted in ascending order, showing performance differences for none, fixed and flexible gaps[1]. These plots give a visual indication of the relative quality of the

[1] Precision is the ratio of positive classifications being positive, and recall is the ratio of correct classified positives. Both are widely used measures for data mining and machine learning

classifiers evolved in the different languages. As can be seen for the correlation plot and the recall plots, there do not seem to be any significant differences even though the motifs with wildcards have marginally higher values. In the precision-plot on the other hand, the patterns with no wildcards are clearly better. This should come as no surprise, as introducing wild-cards in a motif should increase the number of false positives.

Figure 2 show correlations for evolved expressions on the test sets for the different families, indexed from 1 to 44. In this plot, points with the same index represent the same protein family. As can be seen, for most families the differences are very small and the results are strongly correlated (calculated Pearson correlations are higher than 0.8 for all possible combinations.) Thus the added expressibility gained from gaps and flexible gaps might seem unimportant for most families. After reviewing the plots in Figure 1 we might be tempted to conclude that the introduction of gaps, fixed or flexible, gives a benefit. For no gaps, fixed gaps and flexible gaps the summed correlation values over all of SwissProt is 27.4, 28.8 and 29.3 respectively. So the biggest overall benefit is gained from adding fixed wildcards.

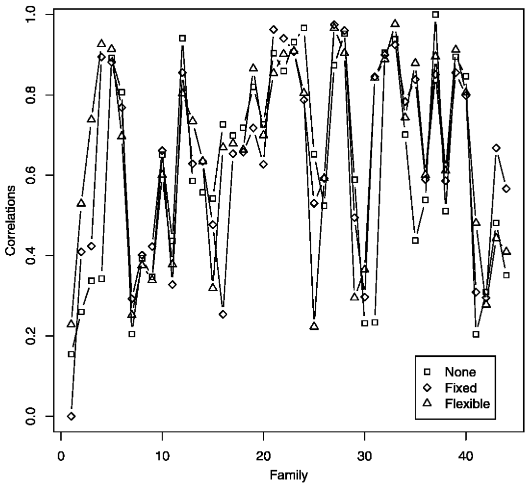


Fig. 2. Correlation coefficients of evolved classifiers on the test sets. Lines are a reading aid, and do not indicate temporal relationships

4.1 Hits and Near Misses

When we compare our results with those of the patterns in the prosite database, we see that on our test-sets we get better precision and recall on the TUBULIN, PA2_ASP and CYTOCHROME_C families. Summing all hits and misses both on the training and testsets, we improve on the Prosite motifs for TUBULIN, PA2_HIS and CYTOCHROME_C. The data for these four different families are

Table 1. Our "hits and near misses." Numbers are rounded to three significant digits, first number is precision and second number is recall

Family	Test	Total	Prosite
TUBULIN	1,1	0.962, 0.997	0.941, 0.986
PA2_HIS	0.963, 0.963	0.964, 0.985	0.946, 0.967
PA2_ASP	0.941, 0.941	0.923, 0.923	0.931, 0.924
CYTOCHROME_C	0.723, 0.991	0.693, 0.993	0.423, 0.989

shown in Table 1. The motif for TUBULIN was found without using wildcards in the terminal sets; the others were found when allowing for fixed size wildcards.

The improved motifs are compared to their Prosite counterparts in table 2. As one might expect, the motifs we find is indeed comparable to those found in the prosite database, and our improvements seem to be only refinements of the original motifs. The CYTOCHROME_C and TUBULIN patterns is actually simplifications of the original patterns. This shows, again, that genetic programming is able to find biological significant motifs.

Table 2. The motifs found by our ListGP algorithm, with comparable performance to those found in the Prosite database. Similar regions of the motifs are printed in bold fonts

Family	ListGP	Prosite
CYTOCHROME_C	C{PC}{PC}CH	C{CPWHF}{CPWR}CH{CFYW}
TUBULIN	GGTG[AS]G	[SAG]GGTG[SA]G
PA2_ASP	D..D.CC.....C	[LIVMA]C{LIVMFYWPCST}CD.....C
PA2_HIS	{NHR}...CC..H{FWP}.C	CC..H..C

5 Concluding Remarks

We demonstrate that ListGP is capable to find motifs with good classification capability, over a range of 44 different families. This is done without any usage of domain specific knowledge, pushing state of the art further and supports the applicability of genetic programming as an alternative tool for biologists. To our knowledge, no researchers in the GP-community have reported experiments on motif discovery in so many protein families before — possibly because searching large populations in large sets of strings is a time-consuming task.

Even though our goal with this paper is not to indulge in a race of improved classifiers for Prosite families, we were able to improve the classification quality for three protein families.

References

1. Halaas A., B Svingen, M Nedland, P Saetrom, Jr. Snove, O., and O.R. Birkeland. A recursive MISD architecture for pattern matching. *IEEE Transactions on Very Large Scale Integraion (VLSI) Systems*, 12(7):727–734, July 2004.

2. Wolfgang Banzhaf, Peter Nordin, Robert E. Keller, and Frank D. Francone. *Genetic Programming – An Introduction; On the Automatic Evolution of Computer Programs and its Applications*. Morgan Kaufmann, dpunkt.verlag, January 1998.
3. A. Brazma, I. Jonassen, I. Eidhammer, and D. Gilbert. Approaches to the automatic discovery of patterns in biosequences. *Journal of Computational Biology*, 5(2):277–304, 1998.
4. Amine Heddad, Markus Brameier, and Robert M. MacCallum. Evolving regular expression-based sequence classifiers for protein nuclear localisation. In Guenther R. Raidl, Stefano Cagnoni, Jurgen Branke, David W. Corne, Rolf Drechsler, Yaochu Jin, Colin Johnson, Penousal Machado, Elena Marchiori, Franz Rothlauf, George D. Smith, and Giovanni Squillero, editors, *Applications of Evolutionary Computing, EvoWorkshops2004: EvoBIO, EvoCOMNET, EvoHOT, EvoIASP, EvoMUSART, EvoSTOC*, volume 3005 of *LNCS*, pages 31–40, Coimbra, Portugal, 5-7 April 2004. Springer Verlag.
5. Yuh-Jyh Hu. Biopattern discovery by genetic programming. In John R. Koza, Wolfgang Banzhaf, Kumar Chellapilla, Kalyanmoy Deb, Marco Dorigo, David B. Fogel, Max H. Garzon, David E. Goldberg, Hitoshi Iba, and Rick Riolo, editors, *Genetic Programming 1998: Proceedings of the Third Annual Conference*, pages 152–157, University of Wisconsin, Madison, Wisconsin, USA, 22-25 July 1998. Morgan Kaufmann.
6. Nicolas Hulo, Christian J. A. Sigrist, Virginie Le Saux, Petra S. Langendijk-Genevaux, Lorenza Bordoli, Alexandre Gattiker, Edouard De Castro, Philipp Bucher, and Amos Bairoch. Recent improvements to the PROSITE database. *Nucl. Acids Res.*, 32(90001):D134–137, 2004.
7. John R. Koza and David Andre. Automatic discovery using genetic programming of an unknown-sized detector of protein motifs containing repeatedly-used subexpressions. In Justinian P. Rosca, editor, *Proceedings of the Workshop on Genetic Programming: From Theory to Real-World Applications*, pages 89–97, Tahoe City, California, USA, 9 July 1995.
8. Brian J. Ross. The evaluation of a stochastic regular motif language for protein sequences. In Lee Spector, Erik D. Goodman, Annie Wu, W. B. Langdon, Hans-Michael Voigt, Mitsuo Gen, Sandip Sen, Marco Dorigo, Shahram Pezeshk, Max H. Garzon, and Edmund Burke, editors, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, pages 120–128, San Francisco, California, USA, 7-11 July 2001. Morgan Kaufmann.
9. Brian J. Ross. The evolution of stochastic regular motifs for protein sequences. *New Generation Computing*, 20(2):187–213, February 2002.

Bayesian Validation of Fuzzy Clustering for Analysis of Yeast Cell Cycle Data*

Kyung-Joong Kim, Si-Ho Yoo, and Sung-Bae Cho

Dept. of Computer Science, Yonsei University
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea
{kjkim,bonanza,sbcho}@cs.yonsei.ac.kr

Abstract. Clustering for the analysis of the gene expression profiles has been used for identifying the functions of the genes and of unknown genes. Since the genes usually belong to multiple functional families, fuzzy clustering methods are more appropriate than the conventional hard clustering methods. However, it is still required to devise natural way to measure the quality of the cluster partitions that are obtained by fuzzy clustering. In this paper, a Bayesian validation method of selecting a fuzzy partition with the largest posterior probability given the dataset is proposed to evaluate the fuzzy partitions effectively. Analysis of yeast cell-cycle data follows to show the usefulness of the proposed method.

1 Introduction

Clustering groups thousands of genes by their similarity of expression levels and helps to analyze gene expression profiles. This organizes the patterns of genes into groups by the similarity of the dataset and has been used for identifying the functions of the genes in the cluster and analyzing the functions of unknown genes. Hard clustering, a hard partitioning method, assigns a sample to only one group. But the real world data like gene expression profiles do not have clear boundaries and they cannot be easily partitioned by hard clustering. Since some genes also belong to multiple functional families, analyzing the genes by hard clustering method has limitations. Fuzzy clustering, unlike the hard clustering, assigns a sample to multiple groups by their grade of membership values [1].

The most important matters that need to be addressed in any clustering method are how many clusters are actually in the dataset and how good the clusters are. Thus, it is necessary to validate each of the fuzzy partition and this evaluation is called cluster validity. Various investigations about these matters have been conducted. Partition coefficient (PC) and partition entropy (CE) were first proposed by Bezdeck [2]. These two cluster validity indexes produce optimal partition at maximum validity measures. Xie-Beni's index (XB) [3] and Fukuyama Sugeno index (FS) [4] are popular in the field of fuzzy clustering. The Xie-Beni index is a ratio of the within cluster sum of squared distances to the product of the number of elements and the minimum between cluster separations, and the Fukuyama Sugeno index measures the compactness and

* This research was supported by the Ubiquitous Computing Research Program funded by the Ministry of Information and Communication of Korea

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 777–784, 2005.

separation of the resulting fuzzy partition after a dataset has been separated into several clusters. However, since the conventional validity indexes are based on the distance between the clusters, we cannot fully represent the structure of the dataset [5].

In this paper, we propose a Bayesian validation method, which evaluates the result of clustering by posterior probability of the fuzzy partitions of given dataset. Unlike the conventional validity indexes, Bayesian validation method never uses the distance between the clusters. It selects the partition with the largest posterior probability in a given dataset. Yeast cell-cycle data is analyzed by the proposed method.

2 Backgrounds

Studies about cluster analysis of the DNA microarray data are summarized in Table 1. Yeung analyzed yeast cell-cycle data by k-means and single-linkage algorithm [6]. Bolshakova and Azuaje used SOM and hard k-means algorithm for clustering and Silhouette index for cluster validation [7]. Also, Eisen analyzed yeast cell-cycle data by fuzzy k-means algorithm and k-means algorithm [8]. Dembele and Kastner used fuzzy c-means algorithm to analyze serum and yeast cell-cycle data [9]. Most of validity indexes used in these researches is all based on the distance between the clusters or between the samples in a cluster: intra-cluster distance and inter-cluster distance.

Table 1. Related works on DNA microarray data

Author	Algorithm	Validity index	Data
Yeung et al. (2001)	K-means Single-linkage	Figure of Merits	Yeast cell-cycle
Bolshakova and Azuaje (2002)	SOM K-means	Dunn's based Index Silhouette Index	Leukemia Lymphoma
Gasch and Eisen (2002)	Fuzzy k-means	N/A	Yeast cell-cycle
Dembele and Kastner (2003)	Fuzzy c-means	Silhouette index	Serum Yeast cell-cycle Human cancer

3 Bayesian Validation Method

All the previous indexes including PC, CE, FS and XB focused on only the compactness and the variation within cluster. However, those indexes lack to provide a correct representation of fuzzy partition in the data since the separation is simply computed by considering only the distance between cluster centroids.

$$\lim_{c \to n} \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} \left\| x_j - v_i \right\|^2 = 0 \tag{1}$$

As shown in Eq (1), if the number of clusters c approaches to the number of samples n, the distance between the cluster centroid and a sample becomes 0. Thus, the traditional indexes lose their ability to validate fuzzy partition for large values of c [5]. Bayesian validation method is a probability-based approach, selecting a fuzzy partition with the largest posterior probability given the dataset. It chooses a partition which has maximum posterior probability given the dataset as an optimal cluster partition. Using Bayes's theorem, the posterior probability given the $Dataset=\{d_1, d_2, \ldots, d_N\}$, could be obtained by multiplication rule and independence rule as follows:

$$P(Cluster \mid Dataset) = \frac{P(Cluster)P(Dataset \mid Cluster)}{P(Dataset)} \tag{2}$$

$$\begin{aligned} P(Cluster \mid Dataset) &= P(Cluster \mid d_1, d_2, ..., d_N) \\ &= P(Cluster \mid d_1) \times P(Cluster \mid d_2) \times ... \times P(Cluster \mid d_N) \end{aligned} \tag{3}$$

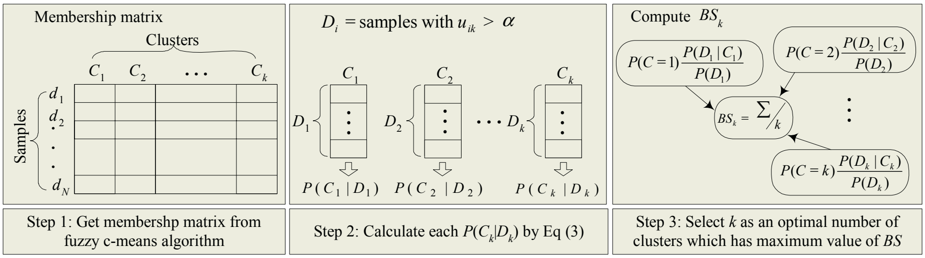
The sum of *P*(*Cluster*|*Dataset*) for all *c* is calculated using Eq (4) and Eq (5) and this value is defined as Bayesian Score (BS). This score indicates how well the fuzzy partition represents the dataset by the posterior probability. Larger value of BS means better cluster partition.

$$\begin{aligned} BS &= \frac{\sum_{i=1}^{c} P(C_i \mid D_i)}{c} = \frac{\sum_{i=1}^{c} P(C_i \mid d_{i1}, d_{i2}, ..., d_{iN})}{c} = \frac{\sum_{i=1}^{c} P(C_i \mid d_{i1}) P(C_i \mid d_{i2}) ... P(C_i \mid d_{iN})}{c} \\ &= \frac{\sum_{i=1}^{c} \prod_{j=1}^{N_i} P(C_i) P(d_{ij} \mid C_i) / P(d_{ij})}{c}, \qquad D_i = \{ d_{ij} \mid u_{ij} > \alpha,\ 1 \le j \le n \},\ N_i = \mathrm{n}(D_i) \end{aligned} \tag{4}$$

In Eq (4), d_{ij} is the *j*th sample which belongs to the *i*th cluster. n(D_i) is the number of D_i's and we select only a sample which has larger membership value (u_{ij}) than certain threshold α for calculation. Since the fuzzy clustering aims mainly to analyze the samples which belong to multiple classes, evaluating the partition with samples whose membership values are larger than certain threshold is more appropriate to group samples by fuzzy clustering method. This threshold is defined as α-cut. Since each membership value u_{ij} represents the belongness of a data x_i to certain cluster *c*, u_{ij} can be substituted for $P(d_{ij}|C_i)$. $P(C_i)$ and $P(d_{ij})$ are calculated as follows:

$$P(C_i) = \sum_{j=1, u_{ij} > \alpha}^{n} u_{ij} \Bigg/ \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}\ ,\ P(d_{ij}) = \sum_{i=1}^{c} P(C_i) P(d_{ij}) = \sum_{i=1}^{c} P(C_i) u_{ij} \tag{5}$$

Figure 1 shows the outline of the proposed method. D_1 includes the samples in cluster C_1 whose membership values are larger than α. Finally, BS is obtained and used to select the optimal fuzzy partition.


Membership matrix
Clusters
C_1 C_2 ... C_k
Samples
d_1
d_2
d_N
D_i = samples with $u_{ik} > \alpha$
C_1 C_2 C_k
D_1 D_2 ... D_k
$P(C_1 | D_1)$ $P(C_2 | D_2)$ $P(C_k | D_k)$
Compute BS_k
$P(C=1)\frac{P(D_1|C_1)}{P(D_1)}$
$P(C=2)\frac{P(D_2|C_2)}{P(D_2)}$
$BS_k = \Sigma / k$
$P(C=k)\frac{P(D_k|C_k)}{P(D_k)}$

Step 1: Get membershp matrix from fuzzy c-means algorithm

Step 2: Calculate each $P(C_k|D_k)$ by Eq (3)

Step 3: Select *k* as an optimal number of clusters which has maximum value of *BS*

Fig. 1. Process of Bayesian validation

The algorithm of Bayesian validation method is as follow:

- Step 1: Compute the membership matrix u_{ij}
- Step 2: Construct D_i by selecting samples ($u_{ij} > \alpha$) in each cluster
- Step 3: Compute $P(D_j|C_j)$, $P(D_j)$, and $P(C_j)$ of D_i
- Step 4: Compute Bayesian Score using the calculated values at step 2
- Step 5: Evaluate the fuzzy partition with the maximum value of BS as optimal one

4 Experimental Results

Yeast cell-cycle data is analyzed with the proposed method. This set contains time-course expression profiles for more than 6000 genes, with 17 time points for each gene taken at 10-min intervals covering nearly two yeast cell cycles (160min). This dataset is very attractive because a large number of genes contained in it are biologically characterized and have been assigned to different phases of the cell cycle. 421 genes are extracted and used for experiments because they are known as informative genes in clustering [10].

Figure 2 shows the results of all the validation methods including the proposed one, where *x* axis represents the number of clusters and *y* axis represents the evaluation value of each validation method. PC and CE have determined the optimal fuzzy partition at *c*=5, FS at *c*=35, XB at *c*=13, and DI at *c*=7 respectively. Unlike the other methods, BS leads to the optimal value at *c*=29. All validity measures show different results and we analyzed biological functions of the cluster partition and its members (genes) which belong to multiple clusters.

We have compared the result of BS which produces the optimal fuzzy partition at *c*=29 with biological knowledge of yeast cell-cycle data [11]. Yeast cell-cycle data represents expression levels of the genes in each of the five cell cycles (Early G_1 – Late G_1 - S - G_2 - M). Each cell cycle includes the genes that show higher expression levels at that cycle time than other cycle times.

By finding clusters that show high peak point in expression levels at certain time in the cycle, we have assigned the cluster to that cycle. Table 2 shows the assigned cluster number and the cycles which they belong to. Clusters that have high expression levels at certain cycle time show low expression level at the other cycle times. Genes assigned between the cycles (intercourse) play a role in regulating the genes that lie in the next cell cycle.

The next step of the analysis is to verify known biological information that the proposed method is indeed able to extract correct information that corresponds to different phases of the yeast cell-cycle data.

Table 3 arranges the genes whose biological functions are known and their cluster number in bracket. Each cycle includes the detailed function groups like DNA replication, biosynthesis, mating pathway and so on. We have confirmed that the results produced by the proposed method are reliable according to the biological knowledge of the genes.

We have chosen special genes whose 1st membership values lie between 0.35 and 0.7, and 2nd membership values are larger than 0.3. These fuzzy genes are belonged to multiple clusters and they provide useful information in gene analysis. Figure 3 shows these fuzzy genes and their biological descriptions with cluster numbers which they belong to. We have classified 4 categories of genes by using the discovered knowledge from Table 7. The genes in cluster 3, cluster 10, cluster 20, and cluster 21 are related to Early G_1 phase. For example, YNL078W belongs to cluster 3 (0.4316) and cluster 19 (0.313888) simultaneously. Actually cluster 3 is related to mating pathway and cluster 19 is related to glycolysis respiration in the same Early G_1 cycle. YNL078W plays multiple roles in Early G_1 cycle. YPR019W, YHR113W, and YHR038W are also fuzzy genes that have multiple functions in cell's life.

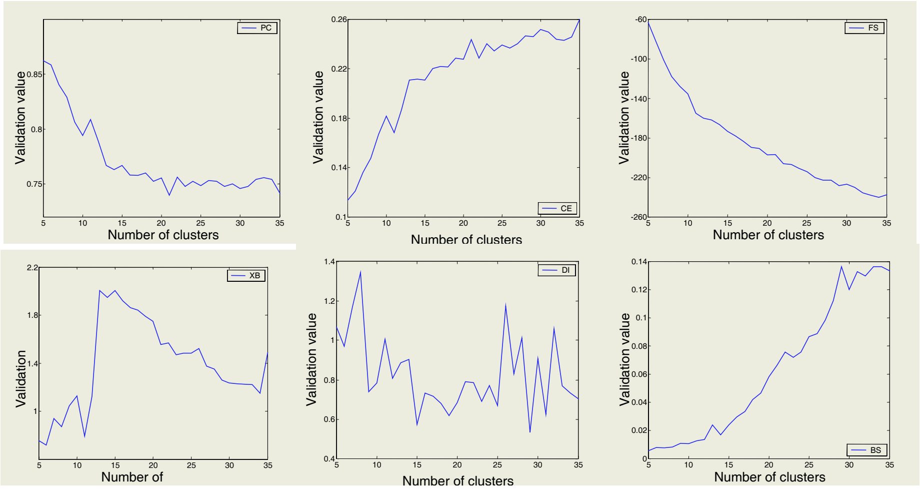


Fig. 2. Preferable values of *c* for yeast cell-cycle data by each cluster validity measure

Table 2. Analysis of cell cycle and clusters

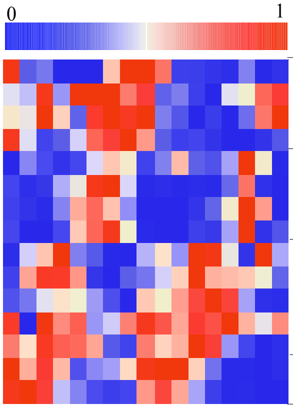
Time (×10 min)	Cell-cycle	Cluster showing peak expression levels on corresponding cycle
0-3	G_1 phase	Cluster5, Cluster6, Cluster4, Cluster24
	intercourse	Cluster2, Cluster12, Cluster26, Cluster28
3-5	S phase	Cluster8, Cluster13, Cluster14, Cluster16
	intercourse	Cluster11
5-7	G_2 phase	Cluster13
	intercourse	Cluster18
7-9	M phase	Cluster7, Cluster17
	intercourse	Cluster10, Cluster21, Cluster3, Cluster20, Cluster19
9-11	G_1 phase	Cluster5, Cluster6, Cluster4, Cluster24
	intercourse	Cluster2, Cluster12, Cluster26, Cluster28
11-13	S phase	Cluster8, Cluster13
	intercourse	Cluster11
13-15	G_2 phase	Cluster0, Cluster13
	intercourse	Cluster18
15-17	M phase	Cluster7, Cluster17

Other fuzzy genes in second category (cluster 12, cluster 24, and cluster 26) are related to Late G_1 phase. Gene like YBR160W, belongs to cluster 12 (0.3982) and cluster 6 (0.3464) simultaneously. Cluster 12 is related to cell cycle regulation and cluster 6 is related to chromosome segregation. Cluster 9, cluster 11, and cluster 13 are related to G_2 phase and cluster 7 and cluster 18 are related to M phase in cell cycle rotation as shown in Figure 3.

We have plotted the fuzzy genes which are analyzed in Figure 3 and their relations are shown in Figure 4. We have used PCA (Principal Component Analysis) to reduce the dimensions of the genes to three and displayed all genes in 3-dimensional space). Fuzzy genes are represented as black cross (X) and rests of genes are represented as different shapes (diamonds, rectangle, triangle, and circle) according to their belonged clusters. As shown in Figure 4, it is clear to see that YHR113W and YHR038W are located between cluster 20 and cluster 21 which are related to Early G_1 phase. Also YHR023 and YOR315W which belong to cluster 7 and cluster 18, are located between these two clusters. These two clusters are related to M phase in cell cycle rotation. Between the other clusters related to Late G_1 phase and G_2 phase, there exist fuzzy genes, providing useful information for further research about unknown genes. Fuzzy genes which have multiple functional families do not have clear boundaries and belong to multiple clusters simultaneously.

Table 3. Analysis of cell cycle and functional groups

Cell-cycle	Functional groups	Genes
Early G_1 phase	DNA replication	YBL023C(10) YEL032W(10) YPR019W(10)
	Mating pathway	YJL157C(3) YKL185W(3)
	Glycolysis, Respiration	YCR005C(20) YCL040W(20) YLR258W(20)
	Biosynthesis	YIL009W(21) YLL040C(21)
Late G_1 phase	Cell cycle regulation	YBR160W(12) YDL127W(12) YGR109C(12) YPR120C(12)
	Chromosome segregation	YDL003W(26) YFL008W(26) YJL074C(26) YKL042W(26) YMR076C(26) YMR078C(26)
	DNA replication	YBR278W(24) YKL045W(24) YLR103C(24) YPR018W(24)
S phase	Chromosome segregation	YDR113C(16) YGR140W(16) YHR172W(16)
	DNA replication	YBL002W(8) YBL003C(8)
	Miscellaneous	YCR035C(14) YER016W(14) YJR137C(14)
G_2 phase	Directional growth	YJL099W(11) YJR076C(11)
	DNA replication	YDR224C(27) YDR225W(27)
M phase	Cell cycle regulation	YGL116W(7) YPR119W(7)
	Transcriptional factor	YDR146C(18) YLR131C(18)
	Directional growth	YCL037C(17)



Gene	Gene description	Clusters
YPR019W	member of the Cdc46p/Mcm2p/Mcm3p family	10, 3
YHR113W	similarity to vacuolar aminopeptidase Ape1p	20, 21
YHR038W	killed in mutagen	20, 21
YNL078W	hypothetical protein	3, 19, 25
YBR160W	"g1,g2" CDC28 cyclin-dependent kinase	6, 12, 24
YDL227C	homothallic switching endonuclease	5, 12, 26
YER070W	ribonucleoside-diphosphate reductase, large subunit	12, 24, 26
YOL017W	similarity to YFR013w	12, 26
YDR464W	regulates spliceosome components	9 ,11
YCR086W	hypothetical protein	9 ,11
YKL052C	hypothetical protein	9 ,11
YPR111W	kinase involved in late nuclear division	9 ,13
YIL050W	cyclin like protein interacting with Pho85p	11, 13
YHR023W	myosin-1 isoform heavy chain	7, 18
YOR315W	hypothetical protein	7, 18

Fig. 3. Analysis of fuzzy genes (gene description and cluster number)

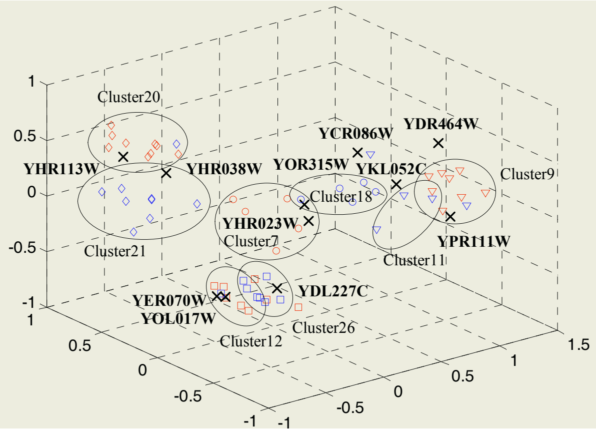


Fig. 4. 3D plot display of fuzzy genes

5 Concluding Remarks

In this paper, a new cluster validation method for the fuzzy partition has been proposed. Bayesian validation method evaluates the fuzzy partition by the posterior probability for the dataset at hand. The best fuzzy partition is obtained by finding the maximum BS value with respect to the number of clusters. We have established α-cut as threshold in computing the value of BS to evaluate various kinds of cluster partitions. We have analyzed the yeast cell-cycle data with the proposed method. To confirm the superiority of the proposed method, the results are verified with biological knowledge.

References

1. A. P. Gasch and M. B. Eisen, Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, vol. 3, no. 11, research 0059.1-0059.22. 2002.
2. J. C. Bezdeck, Cluster validity with fuzzy sets. *J. Cybernit.*, vol. 3, pp. 58-72, 1974.
3. X. L. Xie and G. Beni, A validity measure for fuzzy clustering. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841-846, 1991.
4. Y. Fukuyama and M. Sugeno, A new method of choosing the number of clusters for the fuzzy c-means method. *Proceedings of 5th Fuzzy Systems Symposium*, pp. 247-250, 1989.
5. D. W. Kim, K. H. Lee, D. and H. Lee, Fuzzy cluster validation index based on inter-cluster proximity. *Pattern Recognition Letters*, vol. 24, pp. 2561-2574, 2003.
6. K. Y. Yeung, et al., “Validating clustering for gene expression data,” *Bioinformatics*, vol. 17, no. 4, pp. 309-318, 2001.
7. N. Bolshakova and F. Azuaje, “Cluster validation techniques for genome expression data,” *SIGPRO*, vol. 21, no. 82, pp. 1-9. 2002.
8. A. P. Gasch and M. B. Eisen, “Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering,” *Genome Biology*, vol. 3, no. 11, research 0059.1-0059.22, 2002.

9. D. Dembele and P. Kastner,“Fuzzy c-means method for clustering microarray data,” *Bioinformatics*, vol. 19, no. 8, pp. 973-980, 2003.
10. J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson and P. S. Meltzer, “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks,” *Nature Medicine*, vol. 7, no. 6, pp. 673-679, 2001.
11. R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart and R. W. Davis, “A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, pp. 65-73, 1998.

Rule Generation Using NN and GA for SARS-CoV Cleavage Site Prediction

Yeon-Jin Cho and Hyeoncheol Kim*

Department of Computer Science Education,
Korea University, Seoul, 136-701, Korea
{jx,hkim}@comedu.korea.ac.kr

Abstract. Cleavage site prediction is an important issue in molecular biology. We present a new method that generates prediction rules for SARS-CoV protease cleavage sites. Our method includes rule extraction from a trained neural network and then enhancing the extracted rules by genetic evolution to improve its quality. Experimental results show that the method could generate new rules for cleavage site prediction, which are more general and accurate than consensus patterns.

1 Introduction

Strong interest in automated identification and prediction of cleavage sites have been evoked not only by the huge amount of unprocessed data available but also by the commercial need. The identification and prediction problem is domain-specific, and machine learning methods such as neural networks have therefore been widely used and been successful. In this paper, we present new approaches to rule generation for the cleavage site prediction, and the rule is represented in an explicit form such as “if L@p2 and R@p3, then cleavage”. Prediction of SARS-CoV protease cleavage site was selected as a subject of study.

The first cases of severe acute respiratory syndrome (SARS) were identified in China in November, 2002 and have spread many countries around the world [17]. By late June 2003, the World Health Organization (WHO) has recorded more than 8000 cases of SARS and more than 750 SARS related deaths, and a global alert for the illness was issued due to the severity of the disease [20]. Corona virus (CoV) family consists of four groups as illustrated in figure 1. SARS-CoV is a mutant virus of CoV and belongs to the Group 4. One of the ways to make SARS-CoV incapable is to obstruct the increase of the virus by constraining the activity of proteinase (3CLpro), which is one of the core viruses of SARS-CoV. Therefore, the analysis of CoV cleavage site in the other three groups makes it possible to predict the cleavage site of SARS-CoV and tackle the deseases caused by other upcoming CoV mutant viruses.

SARS-CoV can be incapacitated by restraining the activity of protease (3CL^{pro}), which is one of the main proteins consisting SARS-CoV [11]. Several researches indicate that virus proliferation can be arrested using specific proteinase inhibitors.

Consequently, the protease inhibitors that restrains the virus proliferate can be found by analysing the cleavage site which cleaved by protease, a main protein, in Corona virus. This fact motivated the computational research on the cleavage site

* This work was supported by a Korea University Grant K0400911

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 785–791, 2005.

prediction and analysis. It reduces time and expense of processes of the pathology experiments. The cleavage site analysis enables us to recognize the cleavage candidates and to design the inhibitors of protease [4], [10], [11]. It will assist the cure for SARS and for other diseases caused by Corona viruses.

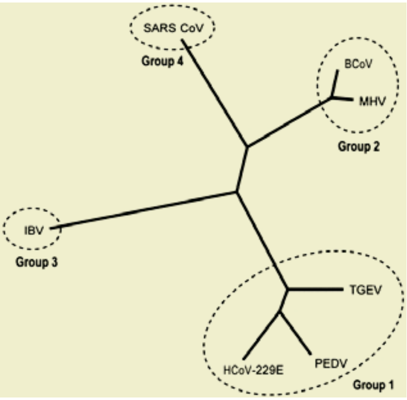


Fig. 1. SARS Corona virus and six other Corona viruses[1]

Machine learning approaches including neural networks have been applied to cleavage site analysis successfully [1], [3], [4], [8], [15]. They however focused on identification or classification of cleavage sites, but not on the explanation of the classification. Prediction knowledge or rule in an explicit format will help us to understand how the sites are classified and provide us with better insights about the specific domain. The idea of prediction rule is not new since consensus patterns and decision trees were used among biologists. However, the consensus patterns are not complete and not accurate enough. Our goal in this paper is to present accurate and robust methods to generate prediction rules of good quality. We used the methods of rule extraction from neural networks and knowledge-based genetic algorithms in this paper.

2 Materials and Methods

In the search for potential inhibitors, important issue is to predict which peptides can be cleaved by SARS-CoV protease. Even limited in the range of an octapeptide, experimental test would be very expensive because the number of possible octapeptides formed from 20 amino acids runs into $20^8 = 2.56\times10^{10}$. Thus, computational methods for cleavage site prediction would be very useful.

2.1 Data Set Preparation

Twenty-four genomic sequences of coronavirus and the annotation information were downloaded from the GenBank database [2], of which 12 are SARS-CoV and 12 are

[1] Marra Ma et al.: The Genome Sequence of the SARS-Associated Coronavirus. SCIENCE VOL 300 (2003)

other groups of coronaviruses. Each sequence of coronavirus genome includes 11 cleavage sites and thus total 264 (= 24×11) sites are available. We eliminated duplicated ones out of the total 264 results and identified final 70 cleavage sites. Each cleavage site of octapeptides includes 8 regions (i.e., 8 positions of P4, P3, P2, P1, P1', P2', P3', P4'). The position p1 is just before the cleavage site; p4 through p1 is N-terminal to the cleavage site and p1' through p4' is C-terminal to the cleavage site. Each region represents one of the 20 amino acids.

For a classification problem such as cleavage site classification, both positive and negative examples are needed. It searches or induces rules that cover positive examples as much as possible and negative examples as little as possible. Negative examples (presumed non-cleavage sites) are created by defining all other Glutamines (Q) in the viral polyproteins as non-cleavable sites [11]. Therefore we obtained 70 positive (i.e., cleavage) and 1267 negative (i.e., non-cleavage) examples in our experimental dataset. Since every site in our dataset has a glutamine (Q) in position P1 (the position just before the cleavage site), the position p1 does not play any role in our classification experiments and thus the symbol “Q@p1” (i.e., amino acid ‘Q’ at position p1) is ignored in the rules generated in our experiments.

For neural network training, each region value that is one of 20 amino acids is converted into 20 binary digits. For example, Alanine(A) among 20 amino acids is represented by 20 bits of 10000000000000000000. Thus, each cleavage sites (i.e., octapeptides) composed of 8 regions is encoded into 160 bits. Class is encoded into either 1 (i.e., cleavage) or 0 (i.e., non-cleavage).

2.2 Methods

Analysis of the Cleavage Sites: Sequence Logo and Decision Tree

Amino acid conservation in multiple sequence alignments may be visualized using sequence logo. Sequence logo is useful for a quick examination of the range in which a sequence signal is present. From the sequence logo in Figure 2, a very strong consensus is evident around the cleavage site. Three consensus patterns from the sequence logo are ‘LQ’, ‘LQ[S/A]’ and ‘[T/S/A]X[L/F]Q[S/A/G]’ [11].

Decision tree is one of the best-known classification techniques in symbolic machine learning. We used C5.0 algorithm and generated the following: “if L@p2 ^ [A/C/G/N/S]@p1’, then cleavage”.

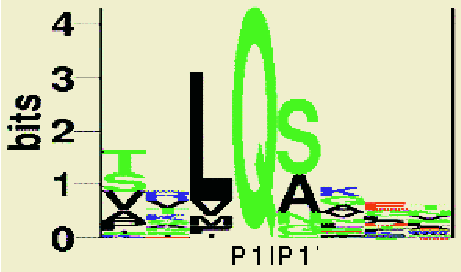


Fig. 2. The sequence logo of SARS-CoV cleavage sites. P1 = N-terminal to cleavage site, P1'= C-terminal to cleavage site

Feed-Forward Neural Network and Rule Extraction

Kiemer, *et al.* used feed-forward neural networks for SARS-CoV cleavage site analysis [11]. They showed that the neural network outperforms three consensus patterns in terms of classification performance. However, they used the neural network for just cleavage site prediction, but not for expressing the sites in explicit knowledge.

There have been many studies for efficient extraction of valid and general rules from a trained neural network [1], [6], [7], [8], [9], [12], [16], [18], [19]. In this paper, we used the OAS (Ordered-Attribute Search) algorithm to extract *if-then* rules from the neural networks [12].

Genetic Algorithm and Knowledge-Based Genetic Algorithm

GA (Genetic algorithm) can be used for searching generalized rules [5], [13], [14]. Individual chromosome in a GA population is a sequence of 8 symbols in which each symbol represents an amino acid or `*' (i.e., don't_care symbol). Then we can say that, for example, the chromosome [**LQS***] represents the rule "If L@p2 and S@p1', then cleavage". We ignore the Q at p1 as mentioned before. Therefore the size of rule space is as huge as 21^8. The GA-based model searches for the best fitted set of chromosomes (i.e., rules) among the 21^8 candidates. One-point crossover is used and crossing point is selected randomly. Mutation occurs on each symbol by 1% and changes its symbol to one of other 20 symbols. Fitness function for a chromosome *n* is defined as follows.

$$f(n) = \frac{nt}{nt + nf + 1} \times 100 + d$$

where *nt (or nf)* is the number of positive (or negative) instances matched by the chromosome rule and *d* is the number of *s (i.e., don't_care symbols) in the chromosome.

The GA-based model generates rules, but the performance is not good enough. The performance was very sensitive to the initial population of chromosomes which was generated by random. Knowledge-based approach to the GA-model is used to restrict the random search space. Domain-knowledge was used as an initial population and GA-model refines and explores from the initial rules. The knowledge-based approach also reduces the GA learning time significantly because it restricts GA search space.

3 Experimental Results

Our experiment includes the following steps; (1) Rule extraction from a trained neural network and comparison of the rules with consensus patterns and decision trees; (2) Rule generation by GA-based model with initial knowledge from consensus patterns, decision tree and neural network rules; (3) Comparison of the rules from neural networks, decision tree and knowledge-based GA. A rule is in the form of "IF condition, THEN class" where class is either of cleavage or non-cleavage. Performance of a rule is evaluated by its coverage and accuracy defined as follows:

$$coverage = \frac{\text{\# of examples matched by the condition part}}{\text{Total \# of examples}}$$

$$accuracy = \frac{\text{\# of true positive examples}}{\text{\# of examples matched by the condition part}}$$

A feed-forward neural network was configured with 160 input nodes, 2 hidden nodes and 1 output node. The neural network was trained and tested by 3-fold cross-validation. Generalization of the neural network is as high as 97.9% while training accuracy is 99.6%. Then we extracted if-then rules from the trained neural network by Kim's OAS algorithm [12], and compared them with consensus rules and decision tree rules.

Next, GA-based model was used with domain knowledge incorporated initially. Domain knowledge was obtained by extracting rules from consensus patterns, decision tree and neural networks. Our experiment shows that the GA rules from neural network knowledge outperform others in terms of the number of quality rules.

Finally, we compare the rule performances between decision tree, neural network and knowledge-based genetic algorithm (KBGA). Table 1 lists the rules with coverage greater than 17% and accuracy greater than 60%. Decision Tree generates rules no better than consensus patterns, while neural network generates four other useful rules in addition to the decision tree rules. The neural network rules were enhanced by GA evolution which was initialized with the neural network rules. The new rule, [S@p4 ^ S@p1'], discovered only by the KBGA is of good quality with accuracy 86.6% and the rule was not discovered by any other methods.

Table 1. The performance of the rules extracted from each classifier algorithm (standard coverage: +17%, standard accuracy: +60%)

	Positive Rules	Coverage(%)	Accuracy(%)
Consensus rules	L@p2 ^ S@p1'	36.35	75.76
	L@p2 ^ A@p1'	26.11	78.26
DT(C5.0) rule	L@p2 ^ S@p1'	36.35	75.76
	L@p2 ^ A@p1'	26.11	78.26
NN rules	L@p2 ^ S@p1'	36.35	75.76
	L@p2 ^ A@p1'	26.11	78.26
	L@p2 ^ E@p3'	21.9	71.43
	V@p4 ^ L@p2	20.71	60.87
	T@p4 ^ L@p2	20.32	77.78
	R@p3 ^ L@p2	17.3	85.71
KBGA rules initialized by NN rules	L@p2 ^ S@p1'	36.35	75.76
	L@p2 ^ A@p1'	26.11	78.26
	L@p2 ^ E@p3'	21.9	71.43
	V@p4 ^ L@p2	20.71	60.87
	T@p4 ^ L@p2	20.32	77.78
	R@p3 ^ L@p2	17.3	85.71
	S@p4 ^ S@p1'	18.73	86.67

4 Conclusion

Prediction or classification rules provide us with explanation about the classification and thus better insights about a domain. We presented a new method that generates rules and improves quality of the rules with the subject of SARS-CoV protease cleavage site prediction. Rules were extracted from a well-trained neural network and then enhanced by genetic evolution. Our experiment presents the rules generated by four different types of approaches:

- Consensus patterns
- Decision Tree
- Neural networks
- Genetic Algorithm initialized by neural network rules

Neural network could generate the rules of high quality that were not discovered by decision trees or consensus patterns. Knowledge-Based Genetic Algorithm (KBGA) model in which the neural network rules were incorporated initially could discover new rules in addition to the neural network rules. The KBGA can be considered as a hybrid model of neural networks and genetic algorithm since knowledge learned by a neural network is enhanced and expanded by GA evolution. The experimental result demonstrates that the hybrid model improves quality of rule generation.

References

1. Andrews, Robert, Diederich, Joachim, Tickle, Alam B.: Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-Based Systems 8(6) (1995) 373-389
2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL.: GenBank: update. Nucleic Acids Res, 32 Database issue: (2004)D23-26
3. Blom N, Hansen J, Blaas D, Brunak S.: Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks. Protein Sci (1996) 5:2203-2216
4. Chen LL, Ou HY, Zhang R, Zhang CT.: ZCURVE-CoV: a new system to recognize protein coding genes in coronavirus genomes, and its applications in analyzing SARS-CoV genomes. SCIENCE DIRECT, BBRC (2003) 382-388
5. De Jong, K.A. and Spears, W.M.: Learning Concept Classification Rules Using Genetic Algorithms. Proceedings of the I Zth. international Conference on Artificial Intelligence (1991) 651-656
6. Fu, LiMin.: Neural Networks in Computer Intelligence. McGraw Hill, Inc (1994)
7. Fu, LiMin.: Rule generation from neural networks. IEEE Transactions on Systems, Man, and Cybernetics 24(8) (1994) 1114-1124
8. Fu, LiMin.: Introduction to knowledge-based neural networks. Knowledge-Based Systems 8(6) (1995) 299-300
9. Fu, LiMin and Kim, Hyeoncheol.: Abstraction and Representation of Hidden Knowledge in an Adapted Neural Network. unpublished, CISE, University of Florida (1994)
10. Gaoa F, Oua HY, Chena LL, Zhenga WX, Zhanga CT.: Prediction of proteinase cleavage sites in polyproteins of coronaviruses and its applications in analyzing SARS-CoV genomes. FEBS Letters 553 (2003) 451-456
11. Kiemer L, Lund O, Brunak S, Blom N.: Coronavirus 3CL-pro proteinase cleavage sites: Possible relevance to SARS virus pathology. BMC Bioinformatics (2004)
12. Kim, Hyeoncheol.: Computationally Efficient Heuristics for If-Then Rule Extraction from Feed-Forward Neural Networks. Lecture Notes in Artificial Intelligence, Vol. 1967 (2000) 170-182
13. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996)
14. Mitchell, M.: An Introduction to Genetic Algorithms. MIT Press (1996)

15. Narayanan, A., Wu, X., Yang, Z.R.: Mining viral protease data to extract cleavage knowledge. bioinformatics, 18(1) (2002) s5-s13.
16. Setino, Rudy, Liu, Huan: Understanding neural networks via rule extraction. Proceedings of the 14th International Conference on Neural Networks. (1) Montreal, Canada (1995) 480-485
17. Stadler K, Masignani V, Eickmann M, Becker S, Abrignani S, Klenk HD, Rappuoli R.: SARS - BEGINNING TO UNDERSTAND A NEW VIRUS. NATURE REVIEWS, MICROBIOLOGY VOLUME 1 (2003) 209-218
18. Taha, Ismali A. and Ghosh, Joydeep: Symbolic interpretation of artificial neural networks. IEEE Transactions on Knowledge and Data Engineering 11(3) (1999) 443-463
19. Towell, Geoffrey G. and Shavlik, Jude W.: Extracting refined rules from knowledgebased neural networks. Machine Learning 13(1) (1993)
20. Yap YL, Zhang XW, Danchin A.: Relationship of SARS-CoV to other pathogenic RNA viruses explored by tetranucleotide usage profiling. BMC Bioinformatics (2003)

A Hybrid Approach to Combine HMM and SVM Methods for the Prediction of the Transmembrane Spanning Region

Min Kyung Kim[1], Chul Hwan Song[2], Seong Joon Yoo[2],
Sang Ho Lee[1], and Hyun Seok Park[1]

[1] Department of Computer Science and Engineering, Ewha University, 11-1 Daehyun-dong
Seodaemun-gu, Seoul, 120-750, Korea
{minkykim,shlee,neo}@ewha.ac.kr
[2] School of Computer Engineering, Sejong University, 98 Gunja, Gwangjin
Seoul, Korea 143-747
peternara@sju.sejong.ac.kr, sjyoo@sejong.ac.kr

Abstract. Transmembrane proteins are the primary targets for the development of new drugs, and a number of algorithms that predict transmembrane topology are publicly available on the Web. In this paper, we present a novel approach using both SVM and HMM methods and we demonstrate that our system outperform the previous systems which only use either HMM methods or SVM methods alone.

1 Introduction

Transmembrane prediction is one of the well-known topics in the bioinformatics field. Since transmembrane protein is a good candidate for discovering new drugs, there arose a natural demand to know novel transmembrane protein, when the genome-sequencing project had been completed. However, although the protein structure is a good resource to predict protein function, the experimental methods to reveal the 3D structure of the transmembrane protein are labor-intensive and time-consuming - the transmembrane region of protein is embedded in the lipid bilayer, and amino acids are composed of an unusually long stretch of hydrophobic residues. This unique feature of transmembrane proteins makes it hard to reveal the structure through the existing techniques such as NMR or X-ray crystallography. For this reason, it is meaningful to predict the structure of transmembrane through machine learning methods.

In this paper, **TMSH**—**T**rans**M**embrane prediction through **S**VM and **H**MM, is introduced by integrating the transmembrane prediction results of the SVM and HMM methods.

2 Related Works

2.1 Transmembrane Region Prediction

Different methods, such as artificial neural network, hidden Markov chains, and the support vector machine, are applied to check the possibility of the existence of a

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 792–798, 2005.

transmembrane region in a protein. The existing transmembrane region prediction tools through machine learning methods are listed in Table 1. As shown in Table 1, most of these tools (ANN, HMM, SVM, etc.) use single kinds of machine learning methods except ENSEMBLE. In the case of ENSEMBLE, though, the disagreement between ANN and two different kinds of the HMM method should be considered once again.

ANN is based on weight matrix, according to their amino acid composition. By using sliding windows, they decide whether there is a transmembrane region. To ensure the reliability of weight matrix, multiple alignments are used. HMM uses a specific architecture for transmembrane proteins. By defining the states of the transmembrane helix residues and those of the inner or outer membrane, and by connecting them in a cycle, they produce a model that closely resembles the transmembrane protein topology. SVM is currently being used for transmembrane prediction, and different coding schemes are being used for this purpose.

Table 1. Transmembrane prediction tools using machine learning approaches

TM Prediction	Machine Learning Methods	Ref
HTP	Artificial neural network	[1]
PHDhtm	Artificial neural network and homology search	[2]
PRED-TMR2	Artificial neural network	[3]
PRED-CLASS	Cascading artificial neural network	[4]
TMHMM 2.0	Hidden Markov model	[5]
HMMTOP	Hidden Markov model	[6]
SVMtm	Support vector machine	[7]
ENSEMBLE	Combination of cascading ANN and HMM	[8]

2.2 Support Vector Machines in Bioinformatics

SVMs are machine learning methods that have only recently been introduced in bioinformatics, and that have shown a good performance in multiple subjects. Protein classification [9], detecting remote homologies [10], and predicting transcription initiation sites [11], are included in this case.

Although the robustness of SVMs with respect to sparse and noisy data is making them the system of choice in a number of applications, they are rarely used for transmembrane prediction. SVMtm is a support vector machine that predicts transmembrane segments [7]. The best-performing coding scheme was selected among the different hydropathy values.

2.3 Problems with Previous Systems

Many transmembrane prediction servers are now available. However, according to the recent TM re-evaluation report which used various measures of accuracies, none of these methods performed consistently well [12]. This means that each approach has its own specialty. It is clear that more general approaches are needed.

ENSEMBLE suggests that the integrated use of different types of machine learning methods (especially, ANN and HMM) outperforms the single methods [8]. For this reason, we decided to use both SVM and HMM architecture for the prediction of transmembrane. As a result, **TMSH** has been developed, integrating these two different methods.

3 Implementation and Evaluation

TMSH has two different modules as shown in Figure 1: SVMs and HMM module. SVMs classify only the possibility of transmembrane by four types of SVM models. The topology information of the protein is acquired by HMM. Training sequence set is required for the construction of supervised learning procedure. The output of query sequences will be labeled as i (inner region of the membrane), M (membrane spanning region), and o (outer region of the membrane).

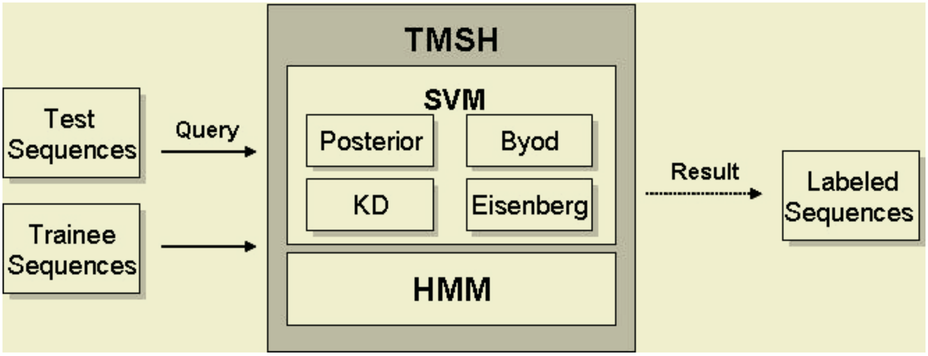


Fig. 1. System architecture of **TMSH**. **TMSH** consists of four different types of SVMs, and of HMM. Their inputs are protein sequences and their outputs are labeled sequences—e.g., "i," "m," and "o"—which stand for inner, transmembrane, and outer regions, respectively

3.1 The SVMs Module

To recognize the transmembrane region, support vector machines (especially LIBSVM) were used, coupled with sliding windows. Input protein sequences were converted into normalized values depending on their hydrophobicity. For example, "0.64", the value of M (amino acids for methionine), was converted into the Eisenberg hydropathy value.

The numeric values of Byod, KD, and Eisenberg in Table 2 came from the results of the experiments performed based on their biochemical properties [13, 14, 15]. The posterior values are the results of the posterior probability of HMM, which is the state *k* probability of the observed sequence *x*. Given a sequence, the posterior probability of each move of the traceback was determined by using the forward-backward algorithm.

$$\begin{aligned} P(x_i, \Pi_i{=}k) &= P(x_1..x_i, \Pi_i{=}k)\, P(x_{i+1}..x_L|\, x_1..x_i, \Pi_i{=}k) \\ &= P(x_1..x_i, \Pi_i{=}k)\, P(x_{\,i+1}..x_L,\ \Pi_i{=}k) \\ &= f_k(i) b_k(i) \end{aligned} \qquad (1)$$

Table 2. The hydrophobicity values which is applied in the four types of SVMs.

Amino Acids	**A**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**
Posterior	0.86	0.92	0.65	0.67	0.87	0.86	0.65	0.90	0.51	0.87
Byod	1.37	1.12	0.17	0.16	1.93	1.03	0.74	2.20	0.19	1.78
KD	0.77	1.00	-1.01	-1.01	1.01	0.03	-0.91	1.67	-1.14	1.44
Eisenberg	0.62	0.29	-0.90	-0.74	1.19	0.48	-0.40	1.38	-1.50	1.06

Amino Acids	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**
Posterior	0.83	0.73	0.76	0.66	0.44	0.84	0.85	0.88	0.82	0.80
Byod	1.39	0.43	0.51	0.35	0.30	0.83	0.89	1.81	-0.14	-0.27
KD	0.80	-1.01	-0.37	-1.01	-1.34	-0.10	-0.07	1.57	-0.14	-0.27
Eisenberg	0.64	-0.78	0.12	-0.85	-2.53	-0.18	-0.05	1.08	0.81	0.26

In the above numerical formula, *x* is the sequence of the training set, and *L* is the length of the sequences. The value of *i* starts from 1 and ends at *L*. The values of *fk(i)* and *bk(i)* are calculated by using the forward-backward algorithm. *P(xi)* is the posterior probability in the specific state *ki*. Thereafter, the normalized value of the posterior probability is sought. Algorithm 1 describes this process (next page).

To determine whether there is a transmembrane region using the SVM method, **TMSH** applied sliding windows, which are widely used in neural networks. The window size 15 was selected from 15-35 based on performance. Each amino acid was converted into four different values, as shown in Table 2. From the i – to the i + window sizes, the value of window size $*$ 2 was acquired, and the maximum value was selected.

Among several types of kernels, **TMSH** uses RBF kernels. The training data set came from the TMHMM training set, which consists of 160 proteins and 63,114 amino acids. The SVM module of **TMSH** was evaluated through cross-validation. The results are shown in Table 3. An Eisenberg hydropathy value showed the best performance.

Algorithm 1 Normalization of the posterior model.

```
char A[j] = {ACDEFGHIKLMNPQRSTVWY}
int N = A.Length ;                              // length of protein
double Amino_Sum[N] ;                       // sum of amino acids
double Amino_Count[N] ;               // number of existing amino acids
double Normal_Amino[N] ;            // normalized values of amino acids

  int i, j = 1;
  for i to L
    if Ki == M                                  // m means transmembrane
      j = 1 ;
      for j to N
        if Xi == A[j]
         Amino_Sum[j] += P(Xi) ;
         Amino_Count[j] ++;
  j = 1;
  For j to N
      Normal_Amino[j] = Amino_Sum[j] / Amino_Count[j] ;
```

Table 3. The results of the SVM modules of **TMSH**

Methods	Correct Location	Single TM Sensitivity	Single TM Specificity
Posterior	48 %	87 %	94 %
Byod	44 %	90 %	90 %
KD	47 %	92 %	84 %
Eisenberg	55 %	92 %	94 %

3.2 The HMM Module

The HMM module of **TMSH** is a part of **PASS** (**TM** module is explained in [16]). It is similar to TMHMM, a well-known transmembrane prediction tool based on the hidden Markov model and part of PASS. Each state has the probability value of the appearance of 20 amino acids, which are trained through the use of Baum-Welch methods. The training data derived from the topology defined positive data that were used in the SVM module. Prediction is carried out through Viterbi calculations that approximate the total probability through the probability of the most likely path among each state.

This method is more accurate than the SVM methods in predicting the correct location (Table 4).

3.3 TMSH: Combining SVM and HMM Modules

TMSH is the integration of the SVM and HMM methods for transmembrane region and topology prediction. The result of **TMSH** is the intersection of the HMM and SVM modules.

Table 4 shows how accurate SVM, HMM, and **TMSH** are, respectively. The correct location was predicted to be less accurate than a single TM's sensitivity and specificity. Among the three methods, **TMSH** was shown to have the best performance. The possibility of falsely predicting the non-transmembrane region and a transmembrane region was eliminated, and the false positive ratio was reduced through a cross-reference between different types of machine leaning methods. **TMSH** confirms the advantage of hybrid of different methods for the detection of the transmembrane segment.

Table 4. Performance test of SVM, HMM and **TMSH**

Methods	Correct Location	Single TM Sensitivity	Single TM Specificity
SVM	55 %	92 %	94 %
HMM	67 %	95 %	96 %
TMSH	75 %	97 %	96 %

4 Conclusion and Future Works

The hybrid approach is commonly used in structure prediction servers. For example, Jpred, which is a consensus secondary structure prediction server, is based on consensus from several methods including DSC, PHD, NNSSP, PREDATOR, ZPRED, and MULPRED [17]. Developers consider the extension of their system such as ANN/HMM/SVM. The hybrid approach of using different learning methods bring to improvement in precision. However, this approach inevitably increases the cost of the prediction procedure.

TMSH integrated the SVM and HMM methods for the prediction of the transmembrane region and topology. **TMSH** is a first attempt to show the advantages of the hybrid system using both SVM and HMM. Although SVM has been successfully applied in multiple areas, the accuracy of SVM for transmembrane region prediction is lower than HMM. There is a possibility that the nature of transmembrane prediction is more adequate to HMM rather than SVM. The SVM module also consists of the four different parts according to their hydropathy values, but they did not show any improvement in performance. We postulate that **TMSH** decreases method-specific false positives and, compensates methods' specific weaknesses by the strength of the alternative technique.

References

1. Fariselli, P., Casadio, R.: HTP: a neural network-based method for predicting the topology of helical transmembrane domains in proteins. *Comput Appl Biosci.* **12** (1996) 41-48.
2. Rost, B., Fariselli, P., Casadio, R.:.Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **5** (1996) 1704-1718.
3. Pasquier, C., Hamodrakas, S.J.: A hierarchical artificial neural network system for the classification of transmembrane proteins. *Protein Eng.* **12**. (1999) 631-634.
4. Pasquier, C., Promponas. V.J., Hamodrakas, S.J.: PRED-CLASS: cascading neural networks for generalized protein classification and genome-wide applications. *Proteins.* **44** (2001) 361-369.
5. Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L.: Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. *J. Mol. Biol.* **305** (2001) 567-580
6. Tusnady, G.E., Simon, I.: The HMMTOP transmembrane topology prediction server. *Bioinformatics.* **17** (2001) 849-850
7. Yuan, Z., Mattick, J.S., Teasdale, R.D.:.SVMtm: support vector machines to predict transmembrane segments. *J Comput Chem.* **25** (2004) 632-636.
8. Martelli, P.L., Fariselli, P., Casadio, R.: An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics.* **19** (2003) i205-i211.
9. Zhang, S.W., Pan, Q., Zhang, H.C., Zhang, Y.L., Wang, H.Y.: Classification of protein quaternary structure with support vector machine. *Bioinformatics.* **19**. (2003) 2390-2396.
10. Saigo, H., Vert, J.P., Ueda, N., Akutsu, T.: Protein homology detection using string alignment kernels. *Bioinformatics.* **20** (2004) 1682-1689.
11. Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T., Muller KR.: Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics.* **16** (2000) 799-807.
12. Chen, C.P., Kernytsky, A., Rost, B.: Transmembrane helix predictions revisited. *Protein Sci.* **11** (2002) 2774-2791.
13. Boyd, D., Schierle, C. and Beckwith, J.: How many membrane proteins are there? *Protein Sci.* **7** (1998) 201-205.

14. Eisenberg, D., Schwarz, E., Komaromy, M. and Wall, R.: Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **179** (1984) 125-142.
15. Kyte. J. and Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157** (1982) 105-132.
16. Kim, M.K., Park, H.S., Park, S.H.: Prediction of plasma membrane spanning region and topology using hidden markov model and artificial neural network. *LNAI.* **3215** (2004) 270-277
17. Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. and Barton, G. J.: Jpred: A Consensus Secondary Structure Prediction Server. *Bioinformatics* **14** (1998) 892-893

Agents in Bio-inspired Computations

V. Kris Murthy

School of Business Information Technology
RMIT University, Melbourne 3000, Victoria, Australia
kris.murthy@rmit.edu.au

Abstract. A multiagent-based programming paradigm (MAP) is described for the evolution of the bio-inspired complex system, e.g., genetic, and active walker (swarm and ant intelligence) models. Since the reaction rules are inherently parallel, any number of actions can be performed cooperatively or competitively among the subsets of the agents, so that the system evolve reaches an equilibrium (or a chaotic or an emergent) state. Practical realisation of this paradigm can be achieved through agent architectures - Adaptive agent and the Java–based Cougaar.

1 Introduction

This paper describes a multiset based 'Multi-Agent programming **p**aradigm (**MAP**)' for bio-inspired computational schemes. These include conventional algorithms, Evolutionary- Genetic algorithms (Michalewicz, and Fogel, [11]), Immunocomputing (de Castro and Timmis [6], Gonchorarova et al.[8],Stepney et al[14] and Active Walker models (ants with scent or multiwalker-paradigm where each walker can influence (repel or attract) the other through a shared landscape based on probabilistic selection, Bonabeau et al [4], Chu et al., [5], Dorigo et al, [7], Kennedy and Eberhart [10], Biomimicry, Pacino [12], Bio-inspired robotics Arkin[1], Bar-Cohen and Breazeal [2].

Principal Features of MAP
The MAP proposed in this paper consists of the following features:

(i) A multiset that contains agents (called the agent-space) whose information is structured in an appropriate way to suit the problem at hand.

(ii) A set of interaction rules that prescribes the context for the applicability of the rules to the agents. Each rule consists of a left-hand side (a pattern or property or attribute) describing the conditions under which the agents can communicate and interact, and a right hand side describes the actions to be performed by the agents, if the rule becomes applicable, based on some deterministic or probabilistic criteria.

(iii) A control strategy that specifies the manner in which the agents will be chosen and interaction rules will be applied, the kinetics of the rule- interference (inhibition, activation, diffusion, chemotaxis) and a way to resolve conflicts when several rules match at once.

(iv) A coordinating agent evaluates the performance of the agents to determine the effectiveness of rule application.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 799–805, 2005.

The rest of this paper is organized as follows: In Sections 2 and 3, general properties and of rule based paradigms are developed. In Section 4 examples of bio-inspired computations realised by MAP are given. Section 5 deals with some agent- tool-kits currently available. Section 6 contains the conclusion.

2 Multi-agent Programming Paradigm

The AOIS (agent oriented information system community) defines an agent as a system that is capable of perceiving events in its environment, or representing information about the current state of affairs and of acting in its environment guided by perceptions and stored information., Woolridge [17].

A multi-agent system can be defined as a loosely coupled network of agents that interact among them and through the environment to solve a problem. Operationally, the multiagent system carries out distributed computation by sending, receiving, handshaking and acknowledging messages and performing some local computations and has the following basic features:

1. An agent can carry out elementary computations and generate random numbers.

2. Initially all agents are inactive except for a specified seeding agent that initiates the computation.

3. Each agent can be active or inactive.

4. An active agent can do local computation, send and receive messages and can spontaneously become inactive.

5. An inactive agent becomes active if and only if it receives a message.

6. Each agent may retain its current belief or revise its belief as a result of receiving a new message by performing a local computation. If it revises its belief, it communicates its revised state of belief to other concerned agents; else it does not revise its solution and remains silent.

3 Kinetics of the Multi-agent System

In order to speed up the use of the multi-agent paradigm we need to consider how to permit multiple agent execution concurrently. This offers the possibility of carrying out parts or all of computations in parallel on distinct processors or performing multiple-simulations simultaneously in a grid or cluster-computing environment. Such possibilities would require the analysis of the rules as to how the rules interfere with There are four ways in which such interference can take place.These interference rules are similar to “Turing’s kinetic rules “ [15], that describe the development of shape, form and pattern in organisms (chemical morphogenesis rules).

1. Enabling Dependence (ED): Agents A(i) and A(j) are called enable dependent (or dataflow dependent) through A(k) if the messages from A (i) creates the required precondition in A(k) and results in a message to A(j) and creates the required precondition in A(j) to act (fire).

2. Inhibit Dependence (ID): Agents A (i) and A (j) are called inhibit dependent, if the actions of A (i) do not create the required precondition in A(k) needed by A (j) and prevents it from executing any action.

3. Opposition Dependence (OD): Agents A(i) and A(j) are opposition dependent (also called data-output dependent) through A(k)), if the order in which A(i) and A(j) enable A(k) and update A(k) produce different results in A(k); that is the objects A(i) and A(j) perform conflicting operations on A(k) and not interleavable. Hence, the local serializability in A(k) is not ensured, if the actions are carried out in different order.

4. Data Antidependence (AD): Agents A (i) and A(j) are data antidependent through A(k) if A(i) enables A(k) and receives the data from A(k) subsequently, the firing of another object A(j) enables A(k) and results in updates of the same set of elements.

Concurrency and Conflicts

We now consider how to speed up the agent system by permitting concurrent transactions between several agents. This would require the analysis as to how the respective internal and external actions interfere with each other when they are applied under varying conditions of context, intention and actions. That is a previous action can create the ***required*** **precondition**. The resulting new action should ***ensure*** that appropriate **post condition** is created after performing the new action. Traditionally, we require that the following two conditions are satisfied for global serialization in distributed computing and transaction processing:

1. At each agent the local schedules the actions are performed in the non-conflicting order **(Local serializability**).

2.At each agent the serialization order of the tasks dictated by every other agent is not violated. That is, for each pair of conflicting actions among actions p and q, an action of p precedes an action of q in any local schedule, if and only if, the preconditions required for p do not conflict with those preconditions required for execution of the action q in the required ordering of **all tasks in all agents (Global serializability).**

The above two conditions require that the preconditions for actions in different agents A(i) and A(j) do not interfere or cause conflicts. In fact it turns out that these conditions are necessary for the stabilization of the multi-agent systems that the computations are locally and globally consistent.

Termination: For termination, the interaction among the agents must come to a halt. When the entire set of agents halt we have an equilibrium state (or a fixed point) also called stability while dealing with exact computation in a deterministic system

Non–termination: This arises when the agents continue to interact indefinitely as in chemical oscillations. Then the multiagent-space reaches a non-equilibrium state.

4 Simulation Examples

MAP can be used for the simulation of bio-inspired computations that include chaotic dynamics, self-organized criticality, and multi-swarming.

(i) Simulating Attractor Dynamics Through Multi-agent Exploration:

The multi agent simulation can speedup the procedure of understanding attractor dynamics by plotting the landscape. We give an example of attractors obtained in the Newton method for solving the complex polynomial f(z) is defined by z = N(z), where N(z) = z – f(z)/ f′(z). Let us consider the computation of the cube root of unity, i.e., solving the equation f (z) = z^3 – 1. Then N (z) = ($2z^3$ + 1)/$3z^2$.

This has attractors (fixed points) at 1, ω = exp 2πi/3, and ω^2 = exp 4πi/3.

The attractor landscape turns out to be very complex for a polynomial with many roots.

Multiple Agents Simulation Procedure: Select a grid of points covering the square region whose corners are at 2 + 2i, 2 – 2i, –2 – 2i, and –2 + 2i. Here we employ four agents. Each agent calculates the value of the 100th iterate of a randomly chosen point and perform the iteration:
z = N(z) and colour the points as follows.
(i) If the distance from the100 th iterate to 1 is less than 1/4, then it is assumed that the point is in the stable set of 1 and the agent colours it blue (in black and white, light gray).
(ii) If the distance from the 100 th iterate to exp 2πi/3 is less than 1/4,then it is assumed that the point is in the stable set of exp 2πi/3 and the agent colours it green (in black and white, dark gray).
(iii) Finally, if the distance from the 100 th iterate to exp 4πi/3 is less than 1/4, then we assume the point is in the stable set of exp 4πi/3 and the agent colours it red (in black and white, black)..
(iv) All the points which are not within 1/4 of one of the roots after 100 iterations of N are left white.

For a 251 by 251 grid the graph produced by the above process is shown in Fig. 1.

Observe that we have three regions — one region defining the stable set of 1, the second one containing the stable set of ω while the third one consisting of the stable set of ω^2.

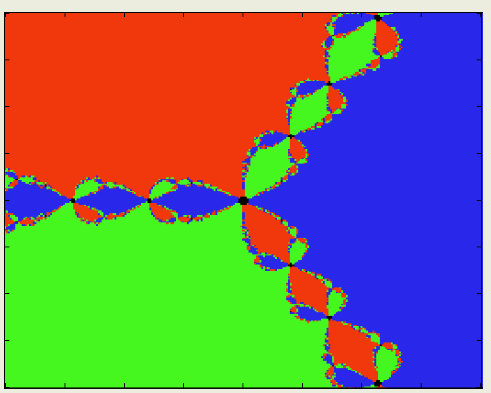


Fig. 1. Attractor sets of Newton iteration z = ($2z^3$ + 1)/($3z^2$) in a 251 × 251 grid

(ii) Multi-agent-Based Group Swarming

MAP is useful for many other types of swarm optimisation in dynamic environment involving particles [10], [4], [16]. Since MAP is based on multisets, Multiswarms can also be simulated [3]. Also MAP can handle Parallel ant colony system simulation suggested by Chu et al. [5].

The Ant Heuristics are based on the model of real ants finding good solutions to the shortest path between their nests and the food sources. Each ant leaves a trail of pheromone, thereby creating a new landscape so that other ants are attracted towards the food source by the scent. The shortest path then turns out to be the path with a maximum density of the scent allowing for evaporation of the scent with time and degradation of the scent with distance. Thus the communication among the ants take place through a shared landscape. Further such ant walks are self-avoiding, that is, each ant avoids revisiting the same location during the search.

The agent model turns out to be quite suitable here, because shared landscape can be simulated through a blackboard and self-avoiding walks can be simulated using a table that stores the locations visited earlier.

To start with the agents initialize their beliefs, by randomly stepping for information. Then with time they update or revise their beliefs through stepping in the environment and communicating with the members of the same group by modifying the landscape further, and then with other groups to obtain collective intelligence and reach an equilibrium state in which shortest path to the food source from the nest is established. The heuristic for a shortest path is given below:

Step 1:

Generate a multiset of agents each agent representing an ant. Let there be G sets in the multiset, each set containing identical agents N(j) for j=0,1, 2,..(G-1). Randomly select an initial node r for each agent. The initial pheromone level on every edge of the graph is set to a small positive quantity f(0) in the blackboard. Set the cycle counter to zero.

Step 2:

Initialization of Beliefs:

Let J (i,j,r) is the set of cities that remain to be visited by the i th agent in the j th group starting from node r ; let f (j,r,s) is the pheromone level between node r and node s, for the j th group. Let v(r,s) = 1/D(r,s) where D(r.s) is the distance between the nodes r and s; let x be a parameter which determines the relative degradation of pheromone level inversely with the distance.

Stepping Rule: Starting from r, calculate the next visited node s, for the i th agent in the j th group as per the rule:

(i) To begin with choose q, a random number in $0 < q < 1$; let q(0) be a constant chosen in the interval between 0 and 1.

(ii) **If** $q < q(0)$, **then**

$$s = \text{Arg Max } u \, \varepsilon \, J(i,j,r) \, [\, f(j,r,u)].\, [v(r,u)\,]^{x} ,$$

else P(i,j,r,s).

Here P(i,j,r,s) is the transition probability from node r to node s for the i th agent in the j th group, given by:

If s ε J (i,j,r) , **then** $f(j,r,s).(v(r,u))^{x} \,/\, \Sigma_{\{u \, \varepsilon \, J(k,\, r)\}} \; [\, f(j,r,u)]\, .[v(r,u)]^{x}$,

else 0;

Updating the Beliefs from Intra-group Knowledge:

Step 3: Local Landscaping Rule:

Update the pheromone level between nodes for each group using:

$f(j,r,s) = (1\text{-}y).f(j,r,s) + y\,\Delta f(r,s)$ where $\Delta f(r,s) = f(0)= 1/\,[nL(n,n)]$

where f(j,r,s) is the pheromone level between nodes r and s for the agents in the j th group. and L(n,n) is the approximate distance of the route between all nodes. Here n is the number of nodes and 0<y<1 is the decay parameter of pheromone.

Continue Stepping and local landscaping until each agent in each group completes the route. In this manner, each agent updates its beliefs and communicates the beliefs to concerned agents.

Step 4:Evaluation:
Calculate the total length travelled by each agent in each group.

Step 5: Global Landscaping Rule:
Let L(j) is the shortest length for the agents in the j th group and z ia pheromone parameter.Update the pheromone level between the nodes for each group using the information on the best available route of the j th group thus:

f(j,r,s) = (1-z).f(j,r,s) +z. Δ f(j, r,s)

where we set

Δ f(j, r,s) = **If** (r,s) belongs to the best route of the j th group,**then** (1/L(j)
else ZERO.

Step 6: Updating Belief from Inter-group Knowledge:
Let L(best) be the best length among the best route of all the groups, i.e., L(best) < L(j), for all j, j = 0,1,2,...,(G-1)

Reassign

f(j,r,s) = f(j,r,s) + w. Δ f(best, r, s)

where:

Δ f(best, r,s)= **if** (r,s) belongs to the best route of all groups **then** 1/ L(best),
else ZERO.

Step 7: Stability:
Perform all steps 2 to 6 **until the system reaches stability,** when **almost all** the agents travel the same route; note that this is a **soft fixed point**.

Simulation Results:
The above algorithm was simulated with a multiset of agents consisting of 5 groups each with 5 agents and the performance studied using the values q(0)= 0.9, x= 2, y=z=w=0.1 and compared to using one group of 25 agents the performance was better. Thus using multi-swarming rather than using a single swarm and setting up a competition within a swarm and competition between different swarms results in a more efficient heuristic. Similar results have been observed by Chu et al [5] and also by Blackwell and Branks[3].

If many different equally likely independent modes m are to searched for, and there are N agents, it seems to be more efficient to choose a multiset each with k elements such that mk = N or k =N/m. If there are correlations between different modes, it will be helpful to use this knowledge with advantage.

5 Multi-agent Toolkits

Shakshuki et al [13] evaluate multiagent tool kits, such as: Java Agent development framework (JADE), Zeus Agent building toolkit (Zeus) and JACK Intelligent Systems. They consider Java support, and performance evaluation. The number of agents they consider is of the order of 32. For the implementation of the paradigm described here, further developments are needed in Agent technology, since we need a very large number of agents to simulate many real-life scientific applications. Gorton et al., [9] have evaluated agent architectures: Adaptive Agent architecture (AAA), Aglets developed by IBM, and the Java based architecture Cougaar. The paradigm described here is well-suited for implementing in Cougaar, a Java based agent architecture, since Cougaar is based on human reasoning. A Cougaar agent consists of a blackboard that facilitates communication and operational modules called plug-in that communicate with one another through the blackboard and contain the logic for the agent's operations. The use of blackboard and direct communication are useful for simulating the problems described in Section 4.

6 Conclusion

This paper described a multiset based Multiagent paradigm (MAP) for the simulation of complex systems. The introduction of probabilistic choices provides for a bio-inspired computational model to study evolutionary biological, chemical and physical systems based on intermittent feedback from the environment.

References

1. Arkin R C,.:Behavior-based Robotics, M.I.T Press, Cambridge, Mass. (1998)
2. Bar-Cohen, Y., and Breazeal, C.: Biologically-Inspired Intelligent Robotics, S.P.I.E. Press, Bellingham, Washington, (2003)
3. Blackwell, T and Branke,J: Multi-swarm Optimization in Dynamic environments, Lecture Notes in Computer Science, 3005,Springer Verlag, New York(2004) 489-500
4. Bonabeau, E., Dorigo,M., and Theraulaz, G. 1999, Swarm Intelligence:From natural to artificial systems, Oxford University Press, Oxford (1999)
5. Chu, S et al.: Parallel Ant colony Systems, Lecture Notes In Artificial Intelligence,, 2871. Springer Verlag, New York(2003)279-284
6. de Castro, L.N and Timmis, J.I.:Artificial Immune Systems: a New computational Intelligence Approach, Springer Verlag, New York(2002)
7. Dorigo,M. et al.: Ant Algorithms, Lecture Notes in Computer Science, 2463, Springer Verlag, New York(2002)
8. Goncharova, L. et al.: Biomolecular Immunocomputing, Lecture Notes in Computer Science,2787, Springer Verlag, New York(2003),102-110
9. Gorton,I, et al.: Evaluating agent Architectures: Cougaar, Aglets and AAA, Lecture Notes in Computer Science, 2940, Springer Verlag, New York (2004)264-274,
10. Kennedy,J. and Eberhart,R.C.: Swarm Intelligence, Morgan Kauffman. London (2001)
11. Michalewicz, Z., and Fogel, D.B.: How to Solve it: Modern Heuristics, Springer Verlag, New York (2000)
12. Pacino,K.M.: Biomimicry of bacterial foraging for distributed optimisation and control, IEEE Control System Magazine,.22(3) (2002)52-68.
13. Shakshuki,E and Jun,Y,.: Multi-agent development toolkits: An Evaluation, Lecture notes in Artficial intelligence,, 3029, SpringerVerlag, New York(2004)209-218
14. Stepney,S, et al,.:Artificial Immune System and the grand challenges for non-classical computation, Lecture notes in Computer Science, 2787, Springer Verlag, New York (2003)204-216
15. Turing, A.M.:The chemical basis for morphogenesis, Phil.Trans. Roy.Soc. London, 237 (1952) 37-79
16. Wolfram, S..:A New kind of Science, Wolfram Media Inc., Champaign, Ill (2002)
17. Woolridge, M,: Introduction to Multi-Agent systems, John Wiley, New York (2002)

Altruistic Punishment, Social Structure and the Enforcement of Social Norms

David Newth

CSIRO Centre for Complex Systems Science
GPO Box 284 Canberra ACT, Australia
david.newth@csiro.au

Abstract. In many social dilemmas, individuals are each better off acting in their own best interest. Social norms provide a mechanism by which group level order can emerge. Often the enforcement of a social norm requires some altruistic punishment. In this paper I explore the role of social structure in the emergence of group level order, in a variation of n-person Prisoners Dilemma played out on a network. The results from this study show that clustering plays an important role in the formation of cohesive groups.

1 Introduction

As individuals, we are each better off when we make use of a common resource without making any contribution to the maintenance of that resource. However, if every individual acted in this manner, the common resource would be depleted and all individuals would be worse off. Social groups often display a high degree of coordinated behaviour that serves to regulate such conflicts of interest. When this behaviour emerges without the intervention of a central authority, we tend to attribute this behaviour to the existence of social norms [1]. A social norm is said to exist within a given social setting when individuals act in a certain way, and are punished when seen to be not acting in accordance with the norm. Dunbar [2, 3], suggests that social structure, and group size play important roles in the emergence of social norms, and cooperative group behaviour.

Like many complex systems, the behaviour of social systems is governed by the organisation of its underlying elements. In this paper I explore the role that community structure plays in the evolution and enforcement of social norms. The remainder of this paper is organised as follows. Section 2 outlines models of social dilemmas, and the norms game that this study is based upon. Section 3 describes the simulation configuration and study results. Section 4, provides a discussion of results and the wider implications of this study. Finally section 5, provides some closing comments and possible future directions.

2 Models of Social Dilemmas

All social dilemmas are marked by at least one deficient equilibrium [4]. It is deficient in that there is at least one other outcome in which everyone is better

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 806–812, 2005.

off. It is equilibrium in that no one has an incentive to change their behaviour. Prisoners Dilemma is the canonical example of such a social dilemma. Prisoners Dilemma is a 2×2 non-zerosum, non-cooperative game, where "non-zerosum" indicates that the benefits obtained by a player are not necessarily the same as the penalties received by another player, and "non-cooperative" indicates that no per-play communication is permitted between players. In its most basic form, each player has two choices: cooperate or defect. Based on the adopted strategies, each player receives a payoff. Fig 1 shows some typical values used to explore the behaviour of Prisoners Dilemma. The payoff matrix must satisfy the following conditions [5]: (1) Defection always pays more; (2) Mutual cooperation beats mutual defection; and (3) Alternating between strategies doesn't pay. Fig 1 also shows the dynamics of this game, the vertical arrows signify the row player's preferences and horizontal arrows the column player's preferences. As can be seen from this figure, the arrows converge on the mutual defection state, which defines a stable equilibrium.

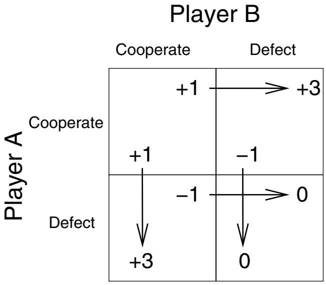


Fig. 1. Prisoners Dilemma. The Payoff structure of Prisoners Dilemma. The game has an unstable equilibrium of mutual cooperation, and a stable equilibrium of mutual defection, this is shown by the arrows, moving away from mutual cooperation to mutual defection

While the 2 person Prisoners Dilemma, has been applied to many real-world situations, there are a number of problems that cannot be modelled. The Tragedy of the Commons is the best known example of such a dilemma [7]. While n-person games are commonly used to study such scenarios, they generally ignore social structure, as players are assumed to be in a well mixed environment [6]. However in real social systems, people interact with small tight cliques with loose long distances connections to other groups. Also traditional n-person games, don't allow player's to punish individuals that do not conform to acceptable group behaviour, which is another common feature of many social systems. To overcome these limitations I will introduce the Norms and Meta-Norms games, which are variations upon the n-person Prisoners Dilemma, which can easily be played out on a network and allows players to punish other players for not cooperating. Fig 2, shows the general structure of the Norms and Meta-Norms games. The following sections describe the games in some detail, and the small world network model, used as a model of social structure.

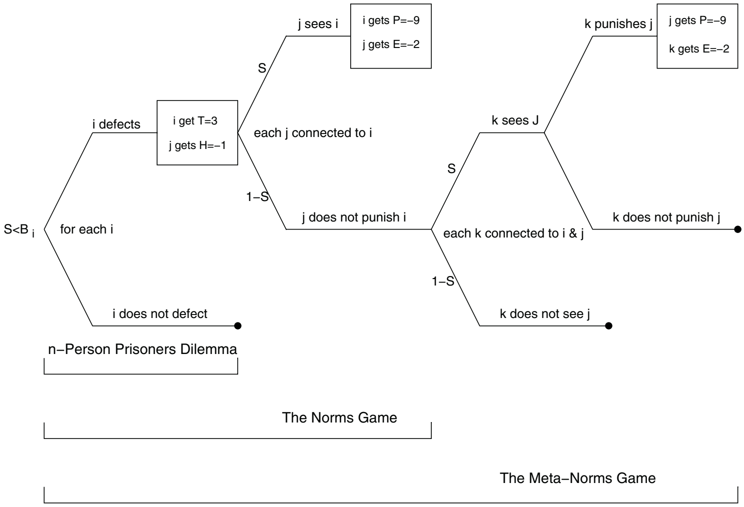


Fig. 2. The Architecture of the Norms and Meta-Norms Games (After Axelrod 1986). Both games start with a variation on the n-person Prisoner's Dilemma. The Norms Game allows, players to punish those players caught defecting. The Meta-Norms Game, allows players to punish those players who do not punish defectors

2.1 The Norm Game

The Norms game and Meta-Norms games [1] are described in Fig 2. It begins when an individual (i) has the opportunity to defect. This opportunity is accompanied by a known chance of being observed defecting (S) by one of i's nearest neighbours. If i defects, he/she gets a payoff T (temptation to defect) of 3, and each other player that is connected to i, receives a payoff H (hurt by the defection) of -1. If the player does not defect then each player receives a payoff of zero. To this point the game is equivalent to an n-person Prisoners Dilemma played upon a network [6]. However should i choose to defect, then one of his n neighbours may see the act (with probability $1-S$), and may choose to punish i. If i is punished he receives a payoff of $P = -9$. However the individual who elects to punish i also incurs an expense associated with dealing out the punishment of $E = -2$. Therefore the enforcement of a social norm –to cooperate– requires an altruistic sacrifice.

From the above description it can be seen that each player's strategy has two dimensions. The first dimension of player i's strategy is *boldness* (B_i), which determines when the player will defect. Defection occurs when $S < B_i$. The second dimension of i's strategy is *vengefulness* (V_i), which is the probability that a player will punish another player if caught defecting. The greater the vengefulness the more likely they are to punish another player.

2.2 The Meta-norm Game

The Meta-Norms game is an extension of the norm game. If player i chooses to defect, and player j elects not to punish i, and i and j have a common neighbour

k, and k observes j not punishing i then k has can punish j. Again j receives the penalty $P = -9$, and like the norm game, k receives the expenses for punishing of $E = -2$.

Like the Prisoners Dilemma, the Norms Game and Meta-Norms Game, have unstable mutual cooperation equilibrium and a stable mutual defection equilibrium. The altruistic punishment, is also an unstable strategy, as punishing an individual also requires a self-sacrifice. The stable strategy is mutual defection with no punishment for defectors. However the global adoption of this strategy means that the population as a whole is worse off, than if the unstable equilibrium strategy of mutual cooperation, with punishment for defectors is adopted.

2.3 Model of Social Structure – Small World Networks

Many complex networks including, ecosystems, neural networks, electrical circuits and even social systems, have been shown to possess a number of common statistical features [8]. Small world networks are typified by two key statistical properties: (1) a higher degree of clustering, than that found in random graph; and (2) an average shortest path length that is approximately the same as that found in random graphs. As a model of small worlds, [9], introduced a one parameter model that simulates many of the characteristics of social networks. The model starts with n nodes (players), connected to their m (for this study $m = 3$) nearest neighbours. Each link is then visited and rewired to point to another node with a probability p. This rewiring procedure introduces long distance connections, reducing path-length characteristics and clustering. This model captures two important aspects of social systems, but it does not capture any system specific properties.

3 Simulations and Results

The two variables (B and V) that make up a strategy, are each allowed to take on a value between $[0, 1]$. The variables represent the probability of defecting and punishing respectively. The variables are each encoded as a 16 bit binary number (as per [1]). The evolution of player's strategies proceeds in following fashion. (1) A small world network of 100 players with a degree of randomness p is created; (2) Each player is seeded with a random strategy; (3) The score or fitness of each play is determined from a given player's strategy and the strategies of the players in their immediate neighbourhood; (4) When the scores of all the players are determined, a weighted roulette wheel selection scheme is used to select the strategies of the players in the next generation. (5) A mutation operator is then applied. Each bit has a 1% chance of being flipped; (6) Steps 3–5 are repeated 500 times, and the final results recorded; (7) Steps 2–6 are repeated $10,000$ times. (8) Steps 1–7 are repeated for p values between 0 and 1 in increments of 0.01. The above experimental configuration was repeated for both the Norms Game; and the Meta-Norms Game. Fig 3, shows the results of these simulations.

From the simulation results we can see that regardless of the social structure, the first order altruistic punishment isn't enough to enforce the social norm of

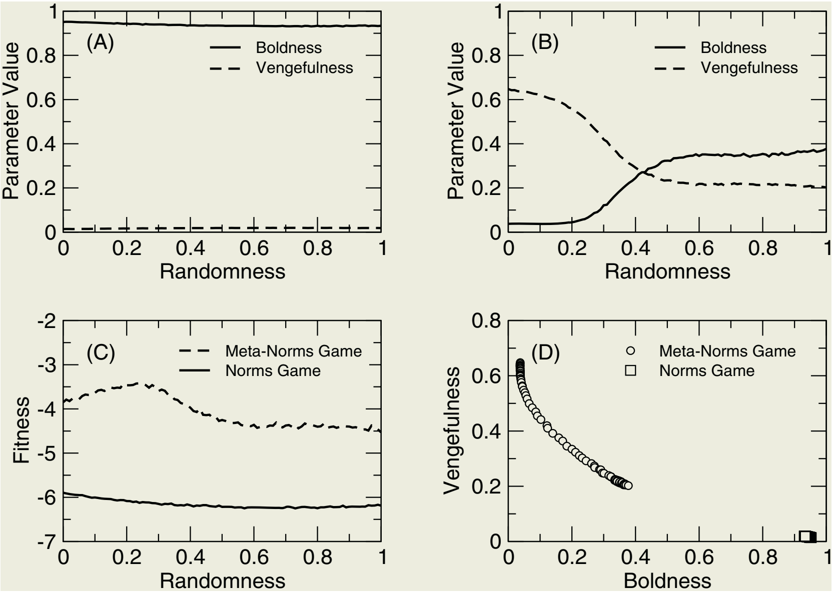


Fig. 3. Simulation Results. (A) Average values for boldness and vengefulness over social networks with varying degrees of randomness for the Norms Game. (B) Average values for boldness and vengefulness over social networks with varying degrees of randomness for the Meta-Norms Game. (C) Comparison of the fitness values for the Norms Game and Meta-Norms Game. (D) Tradeoff between *Boldness* and *Vengefulness*

mutual cooperation. Fig 3(A), shows that regardless of the social structure, the vengefulness decreases to zero, and boldness increases toward one. Essentially all players are attempting to exploit the shared resource, with no fear of being punished. However, for the Meta-Norms game, with second order punishment, there is a distinct set of circumstances when the population as a whole will not exploit the common resource. Fig 3(B), shows that when the social structure is regular, and highly clustered, players boldness decreases, but as the social structure becomes more random (and clustering breaks down), the boldness of a given player increases, and each individual attempts to exploit the common resource. However the level of exploitation is lower than that observed in the Norms game. These differences in system behavior are also seen in the average payoff received by a player (Fig 3(C)). The average payoff per player in the Meta-Norms game is always higher that that received in the Norms game. The average payoff for the Meta-Norms game maximizes just before the transition to a state of global exploitation. Statistical analysis of the network structure reveals that this maximum payoff point coincides with the breakdown of clustering within the network. Finally Fig 3(D), depicts the trade-off between vengefulness and boldness. The Norms game converges to a strategy of low vengefulness and high boldness. While the Meta-Norms game produces a range of behaviors. From the figure it can be seen that there is a trade-off between boldness and vengefulness. The Meta-Norms game produces a wide variety of strategies. These strategies are governed by the topology of the underlying social network. The trade-off surface can be thought of as the set of viable strategies, as nonviable strategies (such as high boldness and vengefulness) are selected against.

4 Discussion

The results from the previous section provide a number of interesting insights into the emergence of social norms and group behaviour. Social structure and second order interactions seem to play an important role in the evolution of group behaviour. In the wider literature, there are many recorded instances where these two factors have been observed to influence group behaviour. In the remainder of this section I will explore three examples:

1. **Animal Innovation.** Japanese Macaque were among the first primates observed by humans to display innovation and diffusion of new novel behaviours to other group members [11]. While many individual animals invent new behaviour patterns, most new behaviours (even if they are beneficial), are unlikely to become fixed within the community. Reader and Laland [10] have shown that there is a link between the social structure of primates and the frequency with which new technologies are uptaken. Populations that tend to be more cliquish are more likely to adopt a new behaviour as members of the clique help to reinforce the novel behaviour.
2. **Social Cohesion.** Dunbar [2], has shown that there is a correlation between neocortex size, and the natural group size of primates. Also correlated with neocortex size, is the cliquishness of the social structure. Dunbar [3] conjectures that the increase in neocortex size, may mean that individuals can manage and maintain more group relationships. The ability to maintain more complex relationships may allow individuals to locally enforce social behaviour. It has also been observed that when primate groups grow too large, social order breaks down, and the troop split into two or more smaller troops, where social order is re-established [3].
3. **Control of Social Behaviour.** The notion of Meta-Norms is widely used in denunciation in communist societies. When authorities accused someone of doing something wrong, others are called upon to denounce the accused. Not participating in this form of punishment is itself taken as a defection against the group and offenders are punished

5 Closing Comments

In this paper I have explored the emergence and enforcement of social norms through the use of two variations on the n-persons Prisoners Dilemma. The simulation results suggest that a combination of second order interactions, altruistic punishment and social structure, can produce coherent social behaviour. Such features have been observed in a number of social systems to enforce norms. The results from this study open a number of interesting future directions: (1) As conjectured by Dunbar [3], social order in primate troops breaks down when the troop becomes too large. This raises the question, what is the relationship between link density, number of nodes and other network statistics, and how do these statistics influence the behaviour of evolutionary games like those described in this paper. (2) Coalitions and factions form and dissolve through time. How do the general results here change if the underlying network is allowed to evolve? (3) Several studies (see [4]) have shown that concepts such as the Nash equilibrium does not hold when human players are substituted for rational players. Do the patterns and tradeoffs described in this paper hold when rational computer players are replaced by human decision makers? All these questions require further experimentation but can be explored in the context of the framework proposed here.

References

1. Axelrod, R. An Evolutionary approach to Norms. Amer. Pol. Sci. Rev. **80**(4). (1986). 1095-1111.
2. Dunbar, R. I. M. Grooming Gossip, and the Evolution of Language. Harvard University Press. (1996).
3. Dunbar, R. I. M. The Social Brain: Mind, Language, and Society in Evolutionary Perspective. Ann. Rev. Anth. **32**. (2003). 163-181.
4. Luce, R. D. and Raiffa, H. Games and Decisions: Introduction and Critical Survey. Dover. New York. (1957).
5. Rapoport, A. Two Person Game Theory. Dover. New York. (1966).
6. Rapoport, A. N-Person Game Theory: Concepts and Applications. Dover. New York. (1970).
7. Hardin, G. The Tragedy of the Commons. Science. **162**. (1968). 1243–1248.
8. Albert, R. and Barabási, A-L. Statistical Mechanics of Complex Networks. Rev. Mod. Phys. **74**. (2002). 47–97.
9. Watts, D. J., and Strogatz, S. Collective dynamics of “small-world” networks. Nature, **393**. (1998). 440–442.
10. Reader, S. M., and Laland, K. N. Animal Innovation. Oxford University Press. (2004).
11. Reader, S. M., and Laland, K. N. Social intelligence, innovation, and enhanced brain size in primates. Proc. Nat. Acad. Sci. **99**(7). (2002). 4432–4441.

WISDOM-II: A Network Centric Model for Warfare

Ang Yang, Hussein A. Abbass, and Ruhul Sarker

Artificial Life and Adaptive Robotics Laboratory (ALAR)
School of Information Technology and Electrical Engineering
University of New South Wales, Australian Defence Force Academy
Canberra, ACT 2600, Australia
{ang.yang,h.abbass,r.sarker}@adfa.edu.au

Abstract. With recognition of warfare as a complex adaptive system, a number of agent based distillation systems for warfare have been developed and adopted to study the dynamics of warfare and gain insight into military operations. These systems have facilitated the analysis and understanding of combat. However these systems are unable to meet the new needs of defence arising from the deeper understanding of warfare and the emergence of the theory of network centric warfare. In this paper, we propose a network centric model which provides a new approach to understand and analyse the dynamics of both platform centric and network centric warfare.

1 Introduction

Traditionally, defence uses human-based warfare simulation to assess risks, optimize missions, and make operational, tactical and strategic decisions. However, this approach is extremely expensive and does not enable analysts to explore all aspects of the problem or repeat simulations.

Recent research [1, 2] shows that warfare is characterized by nonlinear behaviors and that combat is a *complex adaptive system* (CAS). This has attracted attention from researchers and military analysts and a number of agent–based simulation systems [1–4] have been developed to understand and gain insight into military operations. However, with deeper understanding of warfare and the emergence of the theory of network centric warfare (NCW), people have found that it is hard to use these systems to understand and verify the new theory and concepts in warfare, because all existing agent-based distillation systems were built on platform centric warfare and existing agent architectures. In this paper, we propose version II of Warfare Intelligent System for Dynamic Optimization of Missions (WISDOM) [5] which is built on a novel agent architecture, called "network centric multi-agent architecture (NCMAA)". With such agent architecture, WISDOM version II (WISDOM-II) allows people easily to study dynamics in warfare, especially for NCW.

The rest of the paper is organized as follows. We first introduce NCMAA architecture following by the description of WISDOM-II. Then, scenario analysis is conducted. Conclusions are finally drawn.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 813–819, 2005.

2 Network Centric Multi-agent Architecture

NCMAA is purely based on network theory. The system is designed on the concept of networks, where each operational entity in the system is either a network or a part of a network. The engine of the simulation is also designed around the concept of networks.

In this architecture, each type of relationship between the agents defines a network. To design a system, the developer needs first to define an influence diagram of concepts and then develop a finite state machine to control the simulation engine. The influence diagram is a directed graph of concepts defining the interdependency of concepts in the concept space. It provides the basis for establishing a meta-level reasoning system. The finite state machine is a collection of states, each representing the state of a network in the system. The finite state machine represents the sequence of executing each network in the system and the control of the system clock.

NCMAA adopts a two–layer architecture. The top layer, influence network based on the influence diagram, defines the relationship types and how one type of relationship influences other types. Each of these relationship types is reflected in the bottom layer by a set of agents who interact using that relationship.

3 WISDOM Version II

WISDOM-II is developed on NCMAA. It not only uses the spirit of CAS in explaining its dynamics, but also centres its design on fundamental concepts in CAS.

3.1 Architecture

There are 5 concept networks(figure 1) which make up the top layer of the NCMAA in WISDOM-II. Each concept network may have one or more instances which are constituted on blue or/and red agents with their interactions. These instances make up the lower layer of the NCMAA.

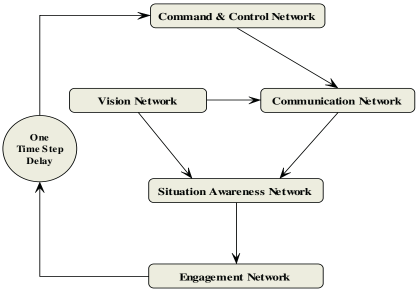


Fig. 1. Influence Network in WISDOM

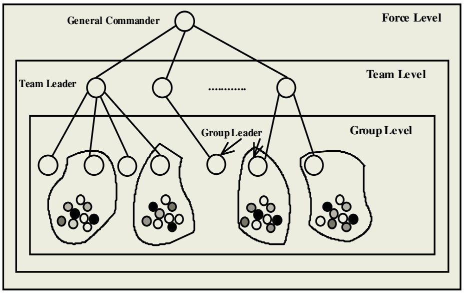


Fig. 2. C2 Hierarchy in WISDOM

- Command & Control (C2) network – defines the command and control hierarchy within one force. Each force has its own commander & control structure. Figure 2 depicts the C2 hierarchy in WISDOM-II. Each group may have a number of agents with different characteristics. We first introduce heterogeneous agents at group level.
- Vision network – a single instance defines the agents which can be seen.
- Communication network – defines the agents which can be communicated with. Communication only occurs within the same force in WISDOM-II. These communication networks could carry two types of information: situation information and commands.
- Situation awareness network – defines the enemy and friend which can be aware of through vision and communication.
- Engagement network – defines the agents being fired at based on the firing agent's situation awareness.

3.2 Agents in WISDOM-II

There are four types of agent in WISDOM-II: combatant agent, group leader, team leader and general commander.

Team leader and general commander are the virtual agents which are sitting in the team bases and force headquarters. The only thing they can do is to communicate with other agents. Combatant agents and group leaders have nine characteristics: health, skill, probability to follow command, visibility, vision, communication, movement and engagement.

Each combatant agent has its own vision which is defined by vision range, detection, correctness function. Communication between agents occurs through a virtual communication channel, which is modelled by noise level, reliability, latency and communication range. Weapon is defined by fire power, fire range and damage radius. WISDOM-II supports point and explosive weapon, and direct and indirect weapon.

Personality is the parameter, called "force", which measures the influence of other agents on the movement of affected agent. Force is a vector quantity which is specified by its magnitude and its direction. Its magnitude is the strength of the influence which is between 0 (no influence) and 1 (strongest influence). The direction of force defines at which direction the influence occurs. There are eight directions, defined by a value between 0 and 1. For each agent, all agents detected or communicated with influence its movement.

3.3 Tactic Decision Making Mechanism

The movement of each agent is determined by its situation awareness and personalities. Only a healthy or wounded agent which is in the battle field can move to a new location. A movement function as in Equation 1 is constructed on the force vectors and an agent moves in the direction of the resultant force vector. The movement algorithm in WISDOM-II is different from that implemented in any other ABD.

$$RF = \frac{\sum_i^n \overrightarrow{F}_i^v + \sum_j^m \overrightarrow{F}_j^c + \overrightarrow{F}^t}{D} \tag{1}$$

RF denotes the resultant force of evaluated agent. n denotes the total number of agents aware of through sensor. $\overrightarrow{F}_i^v$ denotes the influence force from agent i, aware of through sensor. m denotes the total number of agents aware of through communication. $\overrightarrow{F}_j^c$ denotes the influence force from agent j, which is aware of through communication. $\overrightarrow{F}^t$ denotes the influence force from its target. If the evaluated agent is a group member, then the target is its group leader. If the evaluated agent is the group leader, then the target is the group way point. D denotes the distance between the agent and the influencing agent.

3.4 Strategic Decision Making Mechanism

Strategic decision is made by the general commander of each force based on the common operation picture (COP). Three type of decisions can be made for each group based on their missions: advance, defend and withdraw. There are three type of missions in WISDOM-II: defend, occupy and surveillance. The general commander abstracts the whole environment into $n \times n$ (n is predefined by the user.) super cells and calculates the total fire power of the hostile and own force for each super cell, and the total fire power of each group.

For each group with the mission of occupy, the general commander calculates the force power ratio for each of surrounding super cells. The way point will be the super cell with lower hostile fire power ratio. For each group with the mission of surveillance, the way point is the surrounding super cell with highest hostile force power. In order to maximize the information collected, no more than one group will be assigned to the same super cell. No command will be sent to the group with the mission of defend.

The group leader may misunderstand such a command. We introduce two parameters: the probability of misunderstanding and the variance of misunderstanding. The probability of misunderstanding means at which percentage the group leader may misunderstand the command. The variance defines the degree of misunderstand. In other words, it defines the degree of a received way point deviating from its correct location.

3.5 Recovery

One of the most important aspects in military operation is the logistics, where medical treatment system is one of the key components. The model of the artificial hospital is first introduced in WISDOM-II. Each team may have a hospital in the team base, which is defined by the number of doctors and the recovery rate. If the team has a hospital, the wounded agent will move back to the hospital for treatment. Each doctor can treat only one wounded soldier at each time step and the health of that treated soldier will be increased by the recovery rate. If all doctors are already treating, the wounded soldier will be put in the queue to wait for treatment. When the agent is fully recovered, it will move back to the battle field.

3.6 Visualization and Reasoning

One drawback of current ABDs is that they only provide limited information to the analysts during simulation. WISDOM-II fills this gap. Two kinds of information are visualized in WISDOM-II: information about entities in the battle field and information about the interaction between entities.

The architecture of NCMAA makes it possible and easy to conduct real-time reasoning during the simulation. The reason BDI (belief-desire-intention) [6–10] agents do not scale well is because all reasoning is at the individual level. To overcome this, WISDOM-II conducts reasoning at the network (group) level. Two kinds of reasoning are conducted in each time step: time series analysis and correlation analysis.

Time series analysis allows analysts to capture the dynamics over time during the simulation. It includes the damage of each force over time, average degree of each network over time, average path length of each network over time and clustering coefficient of each network over time. When any network collapses, the user may immediately capture it according to these network measures.

Correlation analysis may allow analysts to understand which interaction is playing the key role in damaging their enemy. In each time step, a number of correlation coefficients are calculated, such as correlation coefficient between the damage of their enemy and average degree, average path length of communication network, between the damage of their enemy and situation awareness in agent or force level, and etc.

Based on real-time reasoning, WISDOM-II provides an English-like interface to interpret what happens in the simulation to the user. Following are some examples of output from the reasoning component.

- The red team causes more damage to the blue team. The damage ratio is 1:2.
- With no loss, the blue team causes damage of 16 to the red force.
- The agents in the blue team are coordinating their firing to achieve maximum damage in the red team.
- An average damage of 4 occurred in the red team over the last 5 time steps is probably caused by the activities in the communication network of the blue team and the blue force's situation awareness of enemy on force level.

4 Scenario Analysis

A simple scenario has been built to verify our model. Red force is a traditional force with a large number of soldiers and traditional weapons while blue force is a networked force with a small number of soldiers and advanced weapons. There are two surveillance agents in blue force, which do not have any weapon but they are invisible to red force. The scenario settings for each force are shown in table 1.

Figure 3 presents the damage of each force over time. The damage of red force is much larger while there is much less damage of blue force during the

Table 1. Scenario settings

	Blue Force	Red Force
Number of Agents	11	50
Vision	9 short, 2 long	50 medium
Communication	Networked	medium range
Weapon	2 No weapon, 8 P2P, 1 explosive	50 P2P

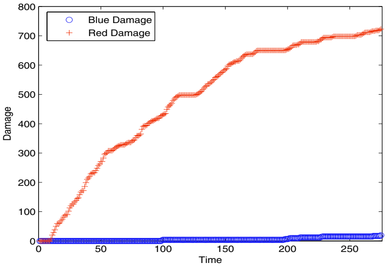


Fig. 3. Damage of each force over time

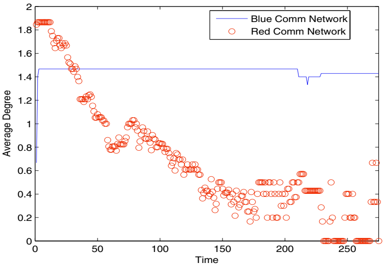


Fig. 4. Average degree of blue and red communication network over time

simulation. Red force has over four times number of agents than blue force. This suggests that communication plays a very important role in combat. Two blue surveillance agents collect information and send them back through communication. Based on COP, the blue agent with powerful weapon, which is long range and explosive weapon, may then shoot their enemies. However, the red agents only have local information. They do not know where their enemy is. Therefore, they cannot win the game. Figure 4 presents the average degree blue and red communication network over time. It supports our intuitive view above. The average degree of the blue communication network is always larger than that of the red communication from about the time step 35. This implies that the information cannot be transmitted efficiently and effectively among red force. They know much less than blue force. So they always get fired upon.

Our scenario demonstrates that fewer number of the networked blue agents can overtake a large number of the red agents. Obviously, one cannot generalize from a single run when using stochastic simulation. However, the objective of this paper is to introduce a new model for warfare and it is not our objective to provide a detailed analysis of warfare simulations.

5 Conclusion

WISDOM-II is much different from other existing ABDs in the sense of architecture, functionality and capability. Since it is developed on NCMAA architecture, it can overcome limitations discussed above. With the help of network measures, WISDOM-II may easily validate the underlying structure. If certain network collapses, the user can immediately detect it through network measures. WISDOM-II conducts real-time reasoning on the network (group) level. Such reasoning captures the domain specific interaction between networks, a natural language interface provides the analyst with online reasoning. Since everything in WISDOM-II is a network or a part of a network, the theory of NCW can be easily modelled, analysed and verified by using WISDOM-II. The use of a network as the representation unit in WISDOM-II also facilitates efficient parallelism based on the network structure, and grounded modelling. In WISDOM-II, a rule based algorithm is used to make strategic decisions, which guides a semi-reactive agent to make tactical decisions. Therefore, the interaction between tactics and strategies is easily captured. The concepts such as information misunderstanding, communication, level of information fusion, etc can now be studied in a unified framework. So WISDOM-II is a promising ABD system. It creates a new approach for analysts to understand the dynamics of and gain insight into warfare.

References

1. Ilachinski, A.: Irreducible semi-autonomous adaptive combat (isaac): An artificial life approach to land combat. Research Memorandum CRM 97-61, Center for Naval Analyses, Alexandria (1997)
2. Lauren, M.K.: Modelling combat using fractals and the statistics of scaling systems. Military Operations research **5** (2000) 47–58
3. Ilachinski, A.: Irreducible semi-autonomous adaptive combat (isaac): An artificial life approach to land combat. Military Operations Research **5** (2000) 29–46
4. Barlow, M., Easton, A.: Crocadile - an open, extensible agent-based distillation engine. Information & Security **8** (2002) 17–51
5. Yang, A., Abbass, H.A., Sarker, R.: Landscape dynamics in multi-agent simulation combat systems. In: Proceedings of 17th Joint Australian Conference on Artificial Intelligence, LNCS, Cairns, Australia, Springer-Verlag (2004)
6. Rao, A.S., Georgeff, M.P.: Bdi agents: From theory to practice. In: Proceedings of the 1st International Conference on Multi-Agent Systems, San Francisco, USA (1995) 312–319
7. Wooldridge, M.J., Jennings, N.R.: Intelligent agents: Theory and practice. Knowledge Engineering Review **10** (1995) 115–152
8. Nwana, H.S.: Software agents: An overview. Knowledge Engineering Review **11** (1996) 205–244
9. Sycara, K.P.: Multiagent systems. AI Magazine **19** (1998) 79–92
10. Wooldridge, M.J.: Intelligent agents. In Weiss, G., ed.: Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence. The MIT Press (1999) 3–44

Adaptation on the Commons

Richard M. Kim[1] and Simon M. Kaplan[2]

[1] School of Information Technology and Electrical Engineering
The University of Queensland, Brisbane 4072, Australia
r.kim@uq.edu.au
[2] Faculty of Information Technology
Queensland University of Technology, 2 George Street, Brisbane 4001, Australia
s.kaplan@qut.edu.au

Abstract. This paper seeks to understand how software systems and organisations co-evolve in practice and how order emerges in the overall environment. Using a metaphor of timetable as a commons, we analyse the introduction of a novel academic scheduling system to demonstrate how Complex Adaptive Systems theory provides insight into the adaptive behaviour of the various actors and how their action is both a response to and a driver of co-evolution within the engagement.

Keywords: complex adaptive systems, information systems, co-evolution, commons

1 Introduction

Over the past 20 years much valuable research has been undertaken to understand Information Systems (IS) engagement, particularly failures or spiraling costs, through the study of critical success factors [1], implementation processes [2, 3], technology resistance [4], or misalignment of the product and organisation [5, 6]. It would be foolish to argue that this significant body of work hasn't advanced our understanding of the IS implementation phenomenon. Yet there exists, by and large, an underlying assumption that regards successful implementation as an isolated feat and not as an ongoing achievement thereby overlooking the contextual and constitutive nature of systems and organisations. One exception to this is the studies of socio-technical systems (STS) that recognise the interdependence of social and technical systems of organisations. Although Kaghan and Bowker [7] question the ongoing usefulness of STS due to its "outdated focus on industrial production and industrial relations" they propose incorporating the rationalist/functionalist approaches of evolutionary economics and complex adaptive systems (CAS) and/or the pragmatist/culturalist approach of actor-network theory as two possible ways of carrying STS forward. Within this broad research agenda, we investigate the incorporation of CAS by regarding implementation as a process of co-evolution where the software system, the vendor, the organisation and its individuals are each forced to continually adapt to the changing context wrought by the movements of one another. Even though the notion that organisations adapt and evolve has become so common that it has assumed the status of a self-evident fact [8], the drivers and mechanics have not been explored in prac-

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 820–826, 2005.

tice. We argue that Complex Adaptive Systems theory [9, 10] provides an excellent lens to study the motor of co-evolution as it seeks to understand the strategies and reinforcement models that drive perturbation within the broader system.

This paper is structured as follows: Section 2 examines how CAS can be applied to enhance our understanding of the co-evolutionary nature of IS engagement. Section 3 re-analyses Kim and Kaplan's [11] case study of the introduction of a knowledge-based intelligent IS intended to fully optimize academic scheduling and resource allocation at a university that we shall refer to as the Australian State University (ASU). Using a metaphor of timetable as commons [12, 13], we demonstrate how CAS provides insight into the adaptive behaviour of the various actors and how their action is both a response to and a driver of co-evolution within the engagement. We discuss these insights in section 4 before concluding.

2 Complex Adaptive Systems

Axelrod & Cohen [9] define a complex adaptive system as a system composed of a population of agents, that we will refer to as actors, that seek to adapt. These actors can equally be actors that act, or artefacts and tools that can be manipulated [14]. Within such a framework, actors or populations of actors interact with their environment and other actors within neighbourhoods and employ a variety of context bound strategies that may be planned and purposeful or conditioned and reactive. Within the system a variety of selection mechanisms also exist that serve to reinforce strategies or actors. A central interest of CAS theorists is how strategies change over time (variation) and how they become more or less common within a population (selection). When selection leads to an improvement it is regarded as adaptation. Co-evolution occurs when populations of actors are continually forced to adapt to the changing context wrought by others' strategies in order to remain relatively fit [15].

Axelrod and Cohen highlight two important subtleties of the CAS framework. Firstly variation need not be successful and secondly local adaptation may actually be to the detriment of the system as a whole. So the question arises, can the diverse goal seeking behaviour of independent actors be harnessed so that local adaptation benefits the overall system rather than harms it?

Kauffman [16] considers this question of co-evolution within a framework of coupled landscapes over which actors adapt and where the actions of one actor cause the landscape to deform thereby acting as input to other actors. This research is largely performed using computational models and as such Anderson [10] warns not to push the landscape metaphor too far when considering organisational settings, but agrees that it is useful for providing insights. One difficulty in transferring these insights into an organisational, and hence social setting, is the interpretive belief that context matters and that situated action is the result of the subjective meaning assigned by actors. We can't objectively categorise actors' strategies as being a success or failure as the concepts of success and failure are parts of the same whole; they depend upon the vantage point from which they are viewed and are rarely final. An alternative would be to adopt an interpretive approach when studying co-evolution and self-organisation and seek to understand the meaning assigned by the actor itself and to track how this changes over time when assessing strategies and understanding reinforcement models.

To this end we re-analyse Kim and Kaplan's [11] interpretative case study of the introduction of a knowledge-based intelligent IS (Stellar Schedule) at the Australian State University (ASU). The research was conducted over 10 months using ethnographic methods to observe 38 meetings and conduct 19 semi-structured interviews. A more detailed account of the research design can be found in [11]. The names of all actors have been changed to maintain anonymity. ASU sought to replace its timetabling system, which could only optimise room allocation against a fixed day/time/staff timetable, with the sophisticated Stellar Schedule which modeled a wide range of constraints to fully optimise day/time/room/staff assignments. Kim and Kaplan demonstrate how order is an emergent property in an IS engagement and one that emerges after socio-technical negotiations have been articulated, contested and resolved. However in their use of actor-network theory solely, they acknowledge difficulties in theorizing the strategies and reinforcement models that drive perturbation within the engagement. The following section demonstrates how CAS can be used to bridge this gap.

3 Analysis of the Case Study

The implementation of an IS occurs within the context of a multiplicity of perspectives as demonstrated by the ASU/Stellar project. We briefly summarise the key groups of actors that have a bearing on the case. Unicited quotations in this section are from interviews with actors of the corresponding group.

Australian State University - ASU is a large research-intensive university that emphasizes excellence in teaching and research that can be summarized by the quote:

> *"As a university we always try to do things that are ahead of the pack."*
> Steering Committee member, interview, 20 August 2004

ASU Management - These actors articulated an opportunistic view of IS implementation that regarded technology as a driver of organisational change and as a shield to deflect the brunt of blame and discontent. Management considered themselves to be "driving the project" and are predominantly inscribed with a view that emphasises rational management and control from the centre by seeking to "maximise physical resource allocation", and to exert control through "standardisation and optimisation".

Central Administrative Staff - For their part the central administrative staff fall within Management's functional hierarchy. Their jobs entail the timely production of the timetable and therefore unsurprisingly they regard the timetable as an end product. Since the central administrative staff work 9am till 5pm Monday to Friday they regard academic's preferences for teaching times as unreasonable; "they should work when they're supposed to!" and "research is what you do when you're not teaching".

Academic Staff - The selection and promotion strategies employed for ASU academics reinforce a culture of research expectation resulting in a perception that "teaching can only hurt you and that it's only through research that one can be promoted". This emphasis on research performance leads academics to be defensive about retaining blocks of "usable" time around their taught engagements.

> *"So what's highly valuable is blocks of uninterrupted time…and so they're right that it can hurt them because it will create what they might see as unusable blocks of time."*
>
> Vice-President (Academic), interview, 26 August 2004

Furthermore, attempts by academics to constrain teaching commitments to the smallest possible number of days per week reinforce central administrators' perceptions as evidenced by the pejorative comments above.

3.1 The Timetabling Commons: Neighbourhoods, Proximity and Interaction

As mentioned earlier we find the metaphor of a commons [12, 13] useful when considering the environment in which this IS engagement occurs. The timetable is a commons that represents a compromise within a conflict-ridden problem space worked out emergently over many years. It allows academics to balance research blocks and teaching blocks, facilitates student access and achieves efficient utilization of physical resources. 'Overgrazing' occurs in the form of increased student enrolment and course offerings and more complex course-conflict rules without a corresponding increase in capacity of physical resources. The selection mechanisms that apply to the different groups of actors reinforce different dimensions of the commons. For academics, time is the common-pool resource and they work together to minimize their interactions with the commons on this basis. Their selection mechanisms, namely tenure and promotion, reward research and not conformance to timetabling processes therefore they adopt various strategies to fence off parts of the commons to maintain 'usable' blocks of time. ASU Management however regards such strategies as being against the 'common good'. Informed by their own selection mechanisms that emphasize optimal utilization of space and maximal student access, they attempt to impose conformance to the 'equitable' allocation of time through the introduction of a new coordination and control device on the commons: the new scheduling system, Stellar Schedule.

In order to understand how systems and organisations co-evolve within this case it is necessary to consider how the key actors, identified in the previous section, interact, how this interaction is facilitated and what strategies are employed during their interaction. Neighbourhoods serve to bring actors together and increase the likelihood of interaction or to keep them apart and act as barriers to interaction. Although some CAS theorists [9] support a dualistic view of neighbourhoods as being physical spaces such as cities, precincts or buildings, or conceptual spaces such as alliances, organisational hierarchies or timezones, we acknowledge these spatial and structural characteristics whilst recognising that neighbourhoods are fundamentally relational.

The organisational hierarchy at ASU was one such neighbourhood which plays an important part in shaping the strategies employed by the actors and in mediating their interaction patterns by creating barriers. These barriers take the form of belonging to different functional reporting lines, such as those of academics and central administrative staff, or being at different levels within the hierarchy.

Under the old system the decision of when classes occurred was made through face-to-face negotiation between the academic and the departmental timetabling coordinator. Although the departmental timetabling coordinator was in most cases an administrative staff member their proximity to, and personal relationships with, the academics led to greater, or sometimes lesser, attempts to accommodate the teaching-constraining strategies of the various academics. In addition to this mutual understanding, the face-to-face nature of the negotiation enabled convergence through fast decision iterations around the fuzzy specification of requirements, i.e. 'I would like two days free per week'. Through the introduction of Stellar Schedule however, Management sought to change these interaction patterns. Rather than decisions being made locally within departments beyond their control, they would be made centrally by Stellar Schedule with data validation performed by the central administrative staff. Although academics were able to provide details of their availability this information was to be filtered through a new semi-permeable barrier erected by ASU Management, namely approval by the relevant Head of Department, in an effort to enforce compliance to new norms of community-minded behaviour. Importantly, although Stellar Schedule optimizes allocation *to* these ranges it does not optimize *within* these ranges. Therefore reporting full availability, as encouraged by Management and Stellar, can result in an individual's timetable being peppered across the commons.

One aspect of interaction that is facilitated by neighbourhoods is the ability for actors to seek to mimic or copy the behaviour of other actors that they perceive to be successful. Complexity theorists call this behaviour signal following and it can be regarded as a strategy to exploit the knowledge or experience of other actors. An alternative strategy is to explore new methods of work or new systems. As a result of the changed interaction patterns imposed by ASU Management, a number of alternative strategies were explored by academic staff in an attempt to maintain stability within their timetable. These strategies can be broadly classified as selfish, reciprocal or altruistic. Under the selfish regime the removal of local interaction led some academics to adopt an extreme opening-gambit position which saw them invert their existing teaching commitments and report this as their unavailability. They then attempted to use the same hierarchy envisaged by administrators to enforce compliance, namely the Head of Department, to sanction their action. Some academics attempted reciprocity in response to the changing landscape by buying coffee for the relevant central administrators in an attempt to forge social bonds that might result in favourable treatment. Whilst others adopted an altruistic approach, reporting full availability with a view to giving the system free reign to choose class times in the best interest of everyone.

For a variety of reasons, reported by Kim and Kaplan but not reproduced here, the central administrators were unable to use Stellar Schedule to produce a fully optimised timetable. Instead it was necessary to roll forward the prior semester's timetable, maintaining days and times, and optimise room allocation solely. Furthermore this occurred on two occasions in consecutive semesters. The effect of this was that the status quo was largely maintained; there was minimal perturbation to the timetable of individual academics. Unaware of the real reasons for the stability, some academics adopting the selfish strategy attributed this 'success' to their own cunning and happily boasted of this to their peers. Likewise those that employed an altruistic approach regard the system with an efficacy greater than it is likely to have when it is finally used. Within this organizational setting therefore, it is the perception of success rather than any objective measure that has created an environment that reinforces or perpetuates these strategies through the signal-following behaviour of other actors.

4 Discussion

We return briefly to a question posed earlier, can the diverse goal-seeking behaviour of independent actors be harnessed so that local adaptation benefits the overall system rather than harms it? Once Stellar Schedule has been adapted sufficiently to enable it to be properly used in the production of the timetable at ASU, Management and the central administrators will be able to enforce compliance to new norms of availability without regard for the academic's selection and reinforcement mechanisms. This has the potential to significantly perturb the research ability of individual academics except for those with a light teaching load. The altruistic behaviour required by Management could actually result in an allocation of time that punishes rather than rewards this behaviour. From the perspective of timetable as a commons the current timetable, which evolved as a collection of compromises and accommodations over many years, creates a commons that contributes to the University achieving its vaunted superior research standing, and gives academics a teaching environment which they generally consider to provide a good compromise between research and teaching time. However, once the understandings governing the commons are removed by fiat, we anticipate that the affected academics are likely to swap to selfish strategies which may be harmful from the perspective of increased student access and community-minded behaviour. Put another way, many academics see the timetable now as a resource to be shared collegially, and are happy to make compromises with one another because they feel they have some control, and that good behaviour can be rewarded. Although a change to selfish strategies may achieve individual stability and look similar in effect in the beginning, the loss of collegiality that would accompany such a shift will have significant implications for the University, analogous to switching from a compromising to adversarial position in a prisoner's dilemma game.

Through the insights of CAS, in particular an appreciation of multiple perspectives and goal seeking behaviour, we come to understand how local adaptation might be steered through changes to reinforcement mechanisms. As noted by Wilson [13] "individual incentives – and, importantly, the willingness to enter into restraining agreements – have to be based on a perception of a beneficial connection between restrained current actions and future states of the natural system." It is unlikely (indeed, undesirable) that ASU modify the primary selection mechanism of research output for academics, so imaginative compromises will be required to convince academics to accept the impact of a widely dispersed timetable. One such mechanism would be further adaptation of Stellar Schedule's optimization algorithm to consider floating blocks of free time - similar to fuzzy specification of availability without creating artificial constraints. This would steer the goal-seeking behaviour of these actors towards positive rather than negative adaptation strategies whilst also satisfying Management's desire for the equitable allocation of common-pool resources on the commons.

5 Conclusion

It is our contention that the majority of IS implementation research centred in the factors, process and politics streams has failed to attend to the co-evolutionary nature of IS engagement. We adopt a stance whereby implementation is a process of co-evolution such that the software system, the vendor, the organisation and its individuals are each forced to continually adapt to the changing context wrought by the movements of one another. By analyzing the introduction of a knowledge-based intelligent IS, we demonstrate how Complex Adaptive Systems can help designers and managers of IS engagements harness and steer the attendant complexities that they encounter. We have seen that actors interact within neighbourhoods using a variety of context-bound strategies to pursue their goals and that this diverse goal seeking behaviour of independent actors could actually be harnessed such that local adaptation benefits the system as a whole. This study is part of a wider work that seeks to turn these insights into an evolutionary information systems development methodology.

References

1. Somers, T.M., Nelson, K.G.: A taxonomy of players and activities across the ERP project life cycle. Inf. & Man., Vol. 41.3. (2004) 257-278
2. Cooper, R.B., Zmud, R.W.: Information Technology Implementation Research: A Technological Diffusion Approach. Mgmt. Sci., Vol. 36.2. (1990) 123-139
3. Kwon, T.H., Zmud, R.W.: Unifying the Fragmented Models of Information Systems Implementation. In: Boland, R.J., Hirschheim, R.A. (eds.): Critical issues in information systems research. Wiley, Chichester (1987) 227-251
4. Markus, M.L.: Power, Politics, and MIS Implementation. Comm. ACM, Vol. 26.6. (1983) 430-444
5. Hong, K.-K., Kim, Y.-G.: The critical success factors for ERP implementation: an organizational fit perspective. Inf. & Man., Vol. 40.1. (2002) 25-40
6. Soh, C., Sia, S.K., Tay-Yap, J.: Cultural fits and misfits: Is ERP a universal solution? Comm. ACM, Vol. 43.4. (2000) 47-51
7. Kaghan, W.N., Bowker, G.C.: Out of machine age?: complexity, sociotechnical systems and actor network theory. J. Eng. Tech. Man., Vol. 18.3-4. (2001) 253-269
8. Morel, B., Ramanujam, R.: Through the Looking Glass of Complexity: The Dynamics of Organizations as Adaptive and Evolving Systems. Org. Sci., Vol. 10.3, Special Issue: Application of Complexity Theory to Organization Science. (1999) 278-293
9. Axelrod, R., Cohen, M.D.: Harnessing complexity: organizational implications of a scientific frontier. Basic Books, New York (2000)
10. Anderson, P.: Complexity Theory and Organization Science. Org. Sci., Vol. 10.3, Special Issue: Application of Complexity Theory to Organization Science. (1999) 216-232
11. Kim, R.M., Kaplan, S.M.: Co-Evolution in Information Systems Engagement: exploration, ambiguity and the emergence of order. In ALOIS*2005 3rd Int Conf on Action in Language, Organisations and Information Systems. Limerick, Ireland, (2005) 166-180
12. Ostrom, E.: Governing the commons: the evolution of institutions for collective action. Cambridge University Press, Cambridge ; Melbourne (1990)
13. Wilson, J.: Scientific Uncertainty, Complex Systems, and the Design of Common-Pool Institutions. In: Weber, E.U. (ed.): The drama of the commons. National Academy Press, Washington, DC (2002) 327-359
14. Kaplan, S., Seebeck, L.: Harnessing Complexity in CSCW. In Seventh European Conference on Computer Supported Cooperative Work (ECSCW 2001). Bonn, Germany, (2001)
15. van Valen, L.: A New Evolutionary Law. Evolutionary Theory, Vol. 1.1. (1973) 1-30
16. Kauffman, S.A.: At home in the universe: the search for the laws of self-organization and complexity. Oxford University Press, New York (1995)

The Emergence of Order in Random Walk Resource Discovery Protocols*

Ricky Robinson and Jadwiga Indulska

School of Information Technology and Electrical Engineering
The University of Queensland
Brisbane, Queensland, Australia
{ricky,jaga}@itee.uq.edu.au

Abstract. While others have attempted to determine, by way of mathematical formulae, optimal resource duplication strategies for random walk protocols, this paper is concerned with studying the emergent effects of dynamic resource propagation and replication. In particular, we show, via modelling and experimentation, that under any given decay (purge) rate the number of nodes that have knowledge of particular resource converges to a fixed point or a limit cycle. We also show that even for high rates of decay - that is, when few nodes have knowledge of a particular resource - the number of hops required to find that resource is small.

1 Introduction

The study contained in this paper presents a mathematical model of random search, and documents the results of experiments performed with the prototype implementation of the Superstring resource discovery protocol [4–6]. The expected results provided by the theoretical model are compared to the results from the experiments.

Section 2 provides a survey of previous work aimed at modelling resource discovery protocols. An overview of the Superstring resource discovery protocol is given in Section 3. In Section 4, a mathematical model of the propagation of resource descriptions through a network is introduced. Results gathered from experiments using the Superstring prototype and their comparison to the behaviour predicted by the mathematical model are presented in Section 5. Conclusions are presented in Section 6.

2 Related Work

A wide variety of resource discovery protocols currently exists. Each utilises a different method for discovering resources on the network. Some resource discov-

* The work reported in this paper has been funded in part by the Co-operative Research Centre for Enterprise Distributed Systems Technology (DSTC) through the Australian Federal Government's CRC Programme (Department of Education, Science and Training)

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 827–833, 2005.

ery protocols utilise random walks, the focus of this paper, to search for resources that match a provided description [2, 6]. These random walk-based protocols can choose to replicate discovered resources (or their descriptions) along the path back to the querier, thereby increasing the chance and speed with which future queries will discover the resource. This can be termed *path replication*, since replication occurs along the return path to the querier. Lv et al. [3] study such protocols, and a variant that uses K simultaneous random walks, in unstructured peer-to-peer networks. Specifically, they study the question "how many copies of each object should there be so that the search overhead for the object is minimized ...?" [3, page 12]. They find that the optimal number of replicas for each resource is proportional to the square root of the relative popularity of the resource (assuming that nodes cannot store replicas of each resource in the system due to storage capacity constraints).

The work reported in this paper describes the replication behaviour of the resource discovery protocol known as Superstring, though the results may apply to other similar protocols. Specifically, we answer the questions *how is knowledge of a resource propagated through a network under path replication and what emergent properties arise from this replication strategy?*

3 Superstring

Superstring is a resource discovery protocol designed to operate in heterogeneous computing environments. It achieves this by defining two underlying routing layers - one for structured, stable environments and one for mobile ad hoc (unstructured) environments - and a single API through which the applications utilise the protocol features.

The structured routing layer utilises a distributed hash table for queries and advertisements. On the other hand, the unstructured routing layer is based on the concept of ant foraging. The unstructured layer is the focus of this paper.

Stigmergy is a method of indirect communication commonly found in nature, which has been thoroughly studied by biologists and complex systems scientists [1, 7, 8]. Ants use stigmergy when foraging for food. After an ant locates a food source, it lays down a pheromone trail on its way back to its nest. Other ants can follow this pheromone trail to the food source. Superstring's unstructured layer query resolution process uses pheromone trails to guide queries to nodes that are likely to be able to resolve those queries.

The following model and experiments make some simplifications to the Superstring unstructured routing layer. Normally, Superstring will prune resource descriptions as they propagate further away from the resource itself, and keep pointers towards the resource, thereby mimicking a pheromone trail and making efficient use of the limited storage on each node. Pruning is not used in this analysis. Instead, entire descriptions are propagated in responses to successful queries. In the world of ants, this is analogous to replicating the entire food source along the path back to the ant nest.

In addition, the model makes several assumptions: (1) that from any node, it is equally probable to forward a message to any other node in the network;

(2) that the number of nodes in the network does not exceed the hop limit for a query; (3) that the network is seeded with a solitary resource, and all queries issued match this single resource, so that the model describes the way in which knowledge of the existence of this single resource spreads through the network; and (4) that queries are issued at a constant interval, and the interval is greater than the time taken for any query to be resolved and returned.

4 A Theoretical Model of Data Distribution

With these simplifications and assumptions in mind, the protocol behaviour can be represented by a mathematical expression. Essentially, the problem is that of calculating a dynamically changing expected value, where the expected value represents the number of nodes in the network that are aware of a particular resource. The nodes that are aware of a resource are said to be *covered* by that resource.

4.1 Coverage Equation

Let χ represent the number of covered nodes. Initially, $\chi = 1$, indicating that only one node has knowledge of the resource to begin with. We distinguish between subsequent values of χ by introducing a subscript, q, so we can rewrite the above as $\chi_0 = 1$.

After q queries, the maximum possible number of nodes that can be visited before the resource is found is equal to the number of uncovered nodes plus one $(n + 1 - \chi_q)$, where n is the number of nodes in the network. The coverage increases as successful queries create pheromone trails, replicating the resource description along the path between the querier and the query resolver. After the first query, more nodes will be aware of the existence of the resource, thereby increasing the probability that the resource will be located in fewer hops for the next query. Therefore, the probability that a particular number of nodes must be visited before the resource is found is dependent upon the current cover.

Differential equations or their discrete analogues, recurrence equations, are often used to model dynamic systems. The recurrence equation shown below provides a way to calculate the expected cover of the network.

$$\chi_{q+1} = \lceil \frac{\chi_q}{n} + \sum_{i=2}^{n+1-\chi_q} i \times (\prod_{k=0}^{i-2} \frac{n - \chi_q - k}{n - k}) \times \frac{\chi_q}{n - i + 1} + \chi_q - 1 - \chi_q \delta \rceil \quad (1)$$

In general, the probability function is constituted by the chance of picking an uncovered node for all except the last hop, in which a covered node is chosen. This is represented by the probability function in Equation 1. The dissipation of the pheromone trail, or in other words, the purging of stale replicated descriptions from nodes along the query route, is represented by the decay factor δ. A decay factor of 0.75 means that during the space of time it takes to issue and resolve four queries, three cache timeouts should occur. Thus, if one query is issued every minute, $\delta = 0.75$ means that the lifetime of a description is eighty seconds. Similarly, a decay factor of 0.25 means that one timeout should occur for every four queries issued, and so on.

The derivation of Equation 1 can be found in [4].

4.2 Model Results

Figures 1 and 2 show the way in which coverage evolves over time, according to the mathematical model, for varying network sizes and varying rates of decay.

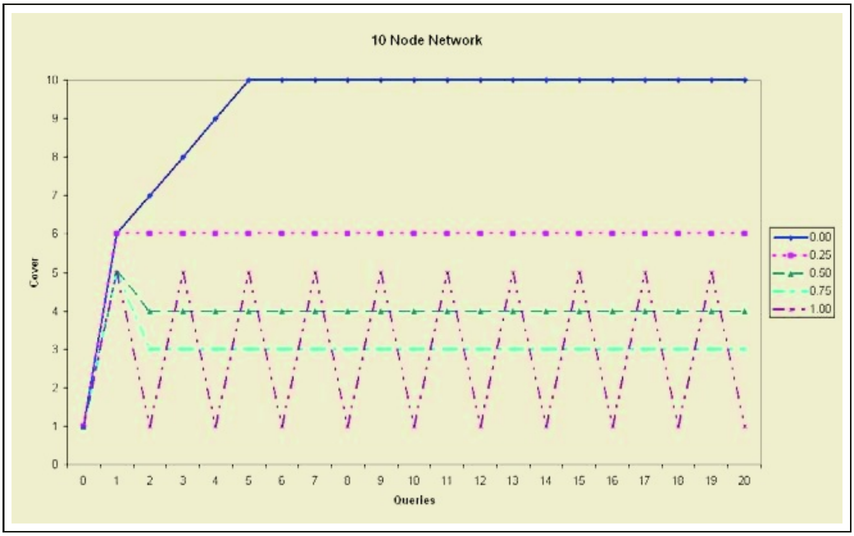


Fig. 1. Change in coverage over time for a 10 node network and varying rates of decay

Each network size and decay pair is associated with a particular cover *attractor*. An attractor is a point or pattern to which a system evolves. There are several kinds of attractors, including point attractors, periodic attractors and strange attractors. Figures 1 and 2 show the attractors for various combinations of network size and decay. For example, in the fifty node network (Figure 2), with $\delta = 0.25$, the system is attracted to a single cover value (13), and is therefore classified as a point attractor. On the other hand, when the decay rate is 1.0 or 0.75, the system settles into a simple cycle between two values, and is therefore properly classified as a periodic point attractor. This shows that order can be achieved in a resource discovery protocol whose foundation lies in a stochastic process. The results for a greater range of network sizes (including networks of up to 100 nodes) can be found in [4].

Expected cover translates into an expected route length, where route length is defined to be the number of nodes traversed en route to a node covered by the resource. Clearly, there is a trade-off between coverage and route length: a

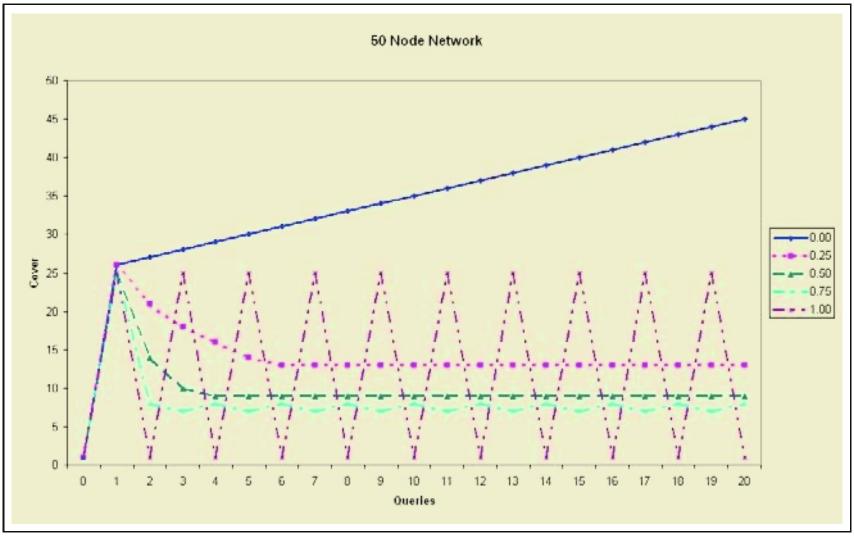


Fig. 2. Change in coverage over time for a 50 node network and varying rates of decay

high coverage leads to shorter route length, while a low coverage leads to longer route length. Figure 3 graphs the attractors for cover and route length with varying rates of decay on a 100 node network. The graph shows that as the rate of decay increases, the expected path length increases much more slowly than the coverage decreases. This suggests that route lengths stay fairly short even for large values of δ.

This analysis, in conjunction with the extended analysis provided in [4], shows that, even in networks on the order of hundreds of devices and where no proactive advertising is performed (an advertisement radius of zero), the number of hops required by a query to discover a matching service is not large, meaning that queries are processed in a timely manner.

5 Experimental Results

The experiments for the local-area protocol were conducted on a stable network of ten nodes. In each experiment, twenty queries were issued at a constant interval of fifteen seconds. Thus, each experiment lasted for approximately five minutes.

Figure 4 depicts the evolution of coverage where resources are purged at each time step (15 seconds) with a probability equal to the decay rate. The results shown in Figure 4 resonate strongly with the results obtained via the mathematical model, shown in Figure 1.

Although node disconnection has not been explicitly modelled or incorporated into the experiments, the idea of node disconnection can, to some extent, be factored into the decay rate. A high rate of disconnection is synonymous with

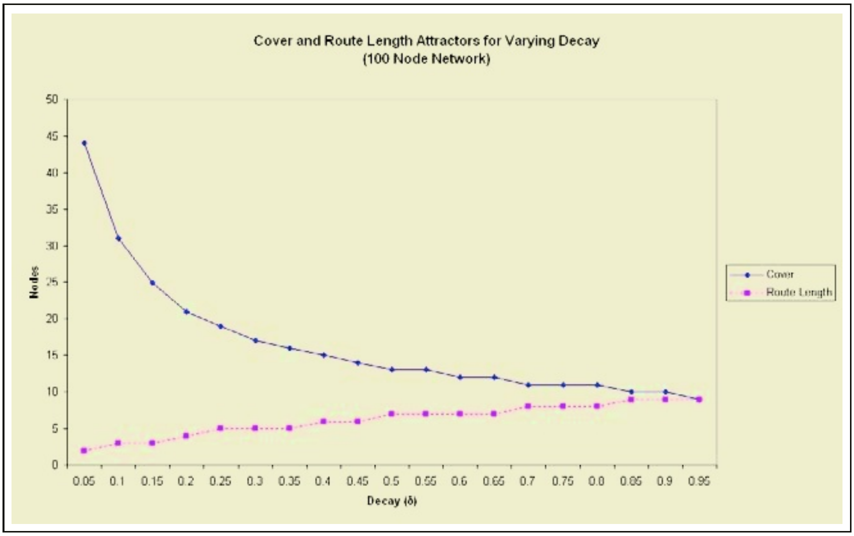


Fig. 3. Cover and route length attractors for increasing decay

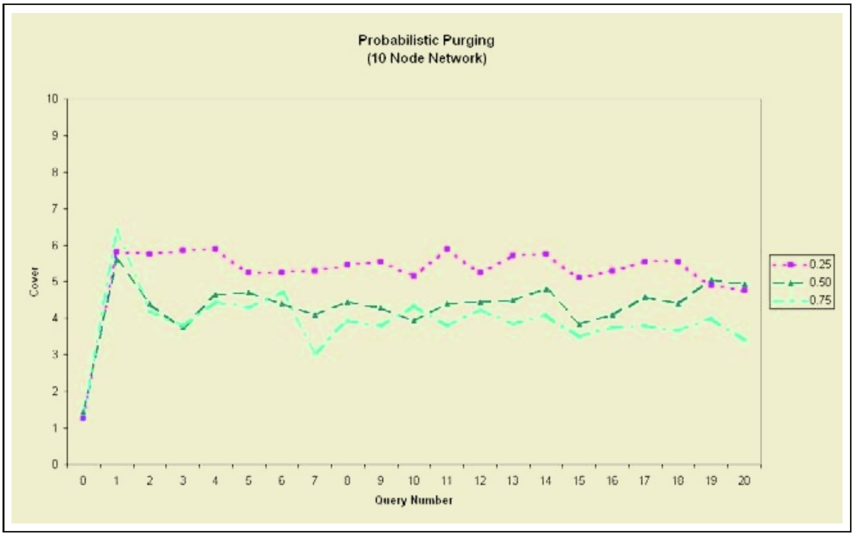


Fig. 4. Prototype results. Resource descriptions were purged with a probability equal to the decay rate in each experiment

a high rate of decay (or short cache lifetimes). In fact, in environments where the rate of disconnection is similar to the rate of node connection (that is where the churn rate is high, but the network size stays fairly constant), the decay rate encapsulates the churn rate.

6 Conclusions

The results presented in this paper show that biological processes, notably stigmergy, are applicable to resource discovery in modern computing environments. Specifically, the natural emergence of order from the stochastic random walk protocol allows the protocol to resolve queries in an efficient manner that is biased towards popular queries. The results also show that a high level of replication is not required to achieve fast and efficient query resolution in random walk protocols.

References

1. J. L. Deneubourg, J. M. Pasteels, and J. C. Verhaeghe. Probabilistic behaviour in ants: A strategy of errors? *Journal of Theoretical Biology*, 105:259–271, 1983.
2. Adriana Iamnitchi, Ian Foster, and Daniel C. Nurmi. A Peer-to-Peer Approach to Resource Discovery in Grid Environments. Technical Report TR-2002-06, Department of Computer Science, University of Chicago and Mathematics and Computer Science Division, Argonne National Laboratory, March 2002.
3. Qin Lv, Pei Cao, Edith Cohen, Kai Li, and Scott Shenker. Search and Replication in Unstructured Peer-to-Peer Networks. In *16th ACM International Conference on Supercomputing(ICS'02)*, pages 84–95, New York, USA, June 2002. ACM Press.
4. Ricky Robinson. *A Resource Discovery Protocol for Modern Computing Environments*. PhD thesis, School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia, March 2005. URL `http://www.itee.uq.edu.au/~ricky/thesis.pdf`.
5. Ricky Robinson and Jadwiga Indulska. Superstring: A scalable service discovery protocol for the wide-area pervasive environment. In *11th IEEE International Conference on Networks*, pages 699–704, Sydney, Australia, September 2003.
6. Ricky Robinson and Jadwiga Indulska. A complex systems approach to service discovery. In *Database and Expert Systems Applications, 15th International Workshop on (DEXA'04)*, pages 657–661, Zaragoza, Spain, August 2004. IEEE Computer Society. ISBN 0-7695-2195-9.
7. David J. T. Sumpter and Madeleine Beekman. From nonlinearity to optimality: pheromone trail foraging by ants. *Animal Behaviour*, 66:273–280, 2003.
8. Guy Theraulaz and Eric Bonabeau. A brief history of stigmergy. *Artificial Life*, 5 (2):97–116, 1999. ISSN 1064-5462.

Supporting Adaptive Learning with High Level Timed Petri Nets

Shang Gao[1], Zili Zhang[1], Jason Wells[1], and Igor Hawryszkiewycz[2]

[1] Deakin University, Geelong VIC 3217, Australia
{shang,zzhang,wells}@deakin.edu.au
[2] University of Technology, Sydney, PO Box 123 Broadway, NSW 2007, Australia
igorh@it.uts.edu.au

Abstract. Supporting adaptive learning is one of the key problems for hypertext-based learning applications. This paper proposed a timed Petri Net based approach that provides adaptation to learning activities by controlling the visualization of hypertext information nodes. Simple examples were given while explaining ways to realize adaptive operations. Future directions were also discussed at the end of this paper.

1 Introduction

Students on Web expect to receive personalized instruction as what they experienced in conventional lectures. Several approaches in this direction are currently under investigation, ranging from learning management systems (IMC, 2005) and (Blackboard, 2005), which focus on course delivery and administrative aspects, and adaptive web-based educational systems (Brusilovsky, 1998)(Henze, 2001) and (Weber, 2001) which offer personalized access and presentation facilities to learning resources for specific application domains (Peter et al, 2004).

Inspired by ideas presented in (Dietrich, 1997)(Frank, 1994)(Rodrigo, 1991) and (Andrew,1994) , a model of hypertext learning state space is proposed (Gao, 2002). Inside of a hypertext learning space, there are many related learning states, each of which consists of a group of knowledge nodes connecting tightly with each other. Transition between states is achieved by calculating states' transition thresholds. A successful transition means all the knowledge nodes of the destination learning state are visible to students. By adjusting the transition thresholds, students' learning paths are controlled.

Given the similarity between learning state space model and Petri Net model, a high level timed Petri Net based approach is proposed to interpret the browsing semantics of a learning state space. By manipulating the time attributes of each state, the adaptation and personalization problem of e-learning activities could be solved.

The paper is organized as follows: an overview of high level timed Petri Net is given as background knowledge in the next section, followed by the description of the proposed timed Petri Net model and its various adaptive operations on the semantic structure of hypertext learning state space. Future research directions will be discussed in the conclusion part.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 834–840, 2005.

2 P-Timed Petri Nets

P-timed Petri Net is an extension of traditional Petri Nets (Tadao, 1989) when used to describe the temporal behavior of a target system. For instance, in a P-timed Petri Net, if one time attribute is associated with place, the firing rules are that a transition is enabled after tokens deposited in its input places take a fixed, finite amount of time. During that time, the tokens are not available. After the time delay, the transition becomes enabled. If fired, tokens are moved into the output places of that transition.

If two time attributes are adopted, one is defined as the minimum delay d^{min} and the other as maximum delay d^{max}, the firing rules are that a transition is enabled after d^{min}; it remains enabled in (d^{min}, d^{max}) interval; if after d^{max}, the enabled transition has not been fired, it is forced to do so, moving tokens from its input places to output places. If the transition can not fire, the token becomes unavailable. This "dead end" should be avoid by setting appropriate (d^{min}, d^{max}) and adjusting them dynamically.

3 P-Timed Petri Net Based Adaptation Model

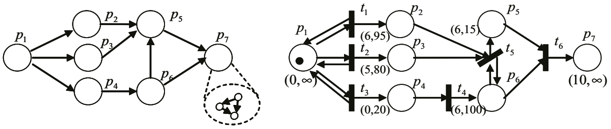
In a hypertext learning state space, each state includes attributes such as contents and numbers of knowledge nodes, recommended learning time, and linkage relation with other states, etc. Different students have different capability, which results in different learning paths and time consumed. To provide personalized learning instructions, a good way is to adjust state attributes to adapt to individual's learning behavior using personal agent, for instance, displaying supplementary materials after assessment, or reducing readings if students are found doing very well. While providing dynamics, the underlying structure should be kept intact. In other words, students are only presented to personalized interfaces with the original underlying structure unchanged.

As discussed above, P-timed Petri Net can be used to model temporal events. If we build a P-timed Petri Net based learning model to control learning activities of students, we can be benefiting from dymanic executive semantics of the Petri Net and consequently obtain a powerful adaptation model.

3.1 P-Timed Petri Net Based Adaptation Model

To build a P-timed Petri Net based adaptation model (P-TPN) for a hypertext learning state space, we first map learning state to place, which is quite straightforward. The second step is to convert outside linkage relation to transitions in P-timed Petri Net model, at the same time remaining the prerequisite constraint and browsing flexibility of hypertext learning state space.

There are several typical structures to consider, for instance, sequence, merging and forking, etc. Due to page limits, detailed transformation will not be discussed here. Interested readers could refer to (Gao, 2005) for further information.



(a) An example space structure

(b) An example P-timed Petri net based model

Fig. 1. An example structure of learning state space and its P-timed Petri Net model

After considering these structures and allocating recommended learning time attributes for each place, a P-timed Petri Net model for a given 7-hypertext learning state space in Figure 1(a) is constructed, as visualized in Figure 1(b). Each of the learning states may include multiple interconnected nodes. Each timed Petri Net place has a delay pair (d_i^{min}, d_i^{max}).

3.2 Firing Rules of P-Timed Petri Net Based Adaptation Model

In a P-TPN model, when a place contains a token, its knowledge node contents are accessible to students. Before d^{min}, the outside link is not click-able or visible. Students need to concentrate on current learning state. A transition t is enabled only after all of its input places p have at least one token and are still within (d^{min}, d^{max}) delay interval. In the (d^{min}, d^{max}) interval, students are free to select visible links and browse other related states. If none of links is selected by d^{max}, an enabled transition will be chosen automatically. The predefined transition priorities usually reflect teachers' teaching preference.

For instance, in Figure 1, place p_1 is visible because of the token deposited in it. Its three links t_1, t_2, and t_3 are immediately click-able due to the 0 value of d_1^{min}. If t_3 is selected, p_1 remains visible and p_4 becomes visible consequently. Student can stay in p_4 for at most 20 time units before forced to leave for p_6. At p_6, students are required to spend at least 6 time units. To activate t_5, students should also access p_2 and p_3 in $(6, 100)$ interval. Only when all the three input places p_2, p_3 and p_6 have tokens and are in their valid intervals, could transition t_5 be enabled. After the smallest delay of d_2^{max}, d_3^{max} and d_6^{max} is taken, t_5 is forced to fire.

If delay pairs are set inappropriately, there might be no overlapping existing in related delay intervals, consequently resulting in "dead end". This problem could be solved by agents which monitor students' status and adjust delay pair values dynamically.

3.3 Timing Control in P-Timed Petri Net Adaptation Model

Timing control for p_i can be easily achieved by setting attributes pair (d_i^{min}, d_i^{max}) particular values, as shown in Table 1.

Different delay pair settings result in different learning activities. Students are required to study for some time defined by d_i^{min}; leaving time defined by d_i^{max} is used to provide flexibility of focusing on detail or leaving. Learning can even be skipped with $(0, 0)$.

Table 1. Place timing control

Timing requirement	Delay pair setting	Example
Normal access time	$d_i^{min} < d_i^{max}$	$(20, 40)$
No maximum constraint	$d_i^{min} \neq d_i^{max}$, $d_i^{max} = \infty$	$(20, \infty)$
No time constraint	$d_i^{min} = 0$, $d_i^{max} = \infty$	$(0, \infty)$
Leaving skipped	$d_i^{min} = d_i^{max} = 0$	$(0, 0)$
Strict access time	$d_i^{min} = d_i^{max} \neq 0$	$(20, 20)$
Learning blocked	$d_i^{min} = d_i^{max} = \infty$	(∞, ∞)

4 Adaptive Operations of P-Timed Petri Net Based Model

Generally, to provide personalized instructions, a P-timed Petri Net model should set attributes of place, transition, net structure, etc. as variables. While students learning, agents running in the background collect information and adjust these attribute values based on a teaching model to adapt dynamics. As discussed above, different learning process can be achieved by setting different time pairs. We now introduce several simple adaptation by this means.

4.1 Content Adaptation

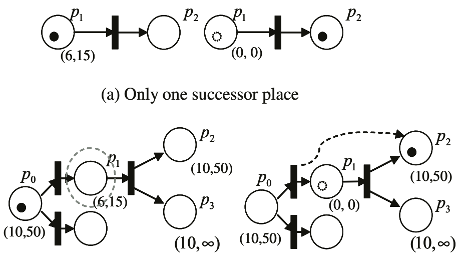
Content adaptation corresponds to displaying or hiding places. In a P-TPN model, a place is skipped by setting its time pair to $(0, 0)$, which is called *learning skip*. When a token flows into the place, no delay is permitted and it enters one of its successors instantaneously. If no such a successor exists, students can try other paths, or accept agents' advise. On all accounts, students are not able to see any contents contained in that place.

Figure 2 (a) illustrates a simply case where the target place p_1 has only one successor place. For a more complex situation like (b), target place p_1 has several successors with different delay pairs. In this case, token would be removed to p_2 because of its smaller d_2^{max} value 50. How to choose an enabled transition depends on the settings of the teaching model or policy. For instance, if faster learning is preferred, the place with smaller d^{min} would be the best candidate. d^{max} is the next variable to consider if all d^{min} are the same, like d_2^{min} and d_3^{min} in case (b).

A place comes to life after changing $(0, 0)$ to its original value pair.

4.2 Link Adaptation

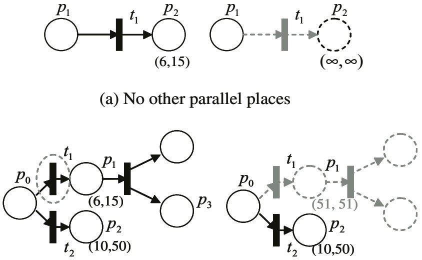
Link adaptation corresponds to deleting or adding transitions. An out-of-date learning state can be removed by adjusting the delay pairs of its output places to (∞, ∞). According to the executive semantics of P-TPN model, when it flows



(a) Only one successor place

(b) Multiple successor places

Fig. 2. Content adaptation



(a) No other parallel places

(b) Multiple parallel places

Fig. 3. Link adaptation

into a place with (∞, ∞) delay, a token could only leave after ∞ time units. Such "dead-end" transition would never be activated, so does its corresponding link.

Without setting the extreme value ∞, a link could also be removed by choosing the largest d_i^{max} of its parallel places and setting its output places' delay pairs to $(d_i^{max} + 1)$.

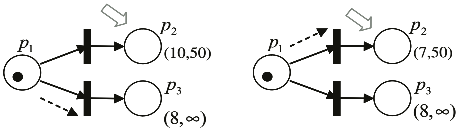
Figure 3 demonstrates how deletion of a transition (or link) can be realized. Case (a) is a simply case where the target transition t_1 has only one output place. For a more complex situation like (b), target transition t_1 has another parallel transition t_2 coming from the same input place p_0. Value (∞, ∞) definitely makes t_1 never be chosen. However, value $(51, 51)$ for p_1 also works given $d_2^{max} = 50$.

A transition would recover after changing the related delay pairs back to its original values.

4.3 Timing and Priority Adaptation

Timing adaptation can be achieved by dynamically adjusting delay pair values. When authoring educational hypertext, timing attributes are set by default. Examination timing could be reasonable, however, predefined timing for some learning state might not be appropriate. For instance, students devote too much time on some details and do not finish browsing before d^{max}. Under such kind of circumstances, adjustment is necessary.

Changing delay pairs affects the transition sequence (or learning path). A place has higher priority if its d^{min} set to the smallest among its parallel places. Its priority degrades when granted a larger d^{min}. Figure 4 illustrates this kind of situation.



(a) Before (b) After

Fig. 4. Priority adaptation

p_3 has a higher priority of being chosen because its $d_3^{min} = 8$ is less than $d_3^{min} = 10$. Things are reversed after d_2^{min} is changed from 10 to 7 which is less than $d_3^{min} = 8$.

5 Discussion and Conclusion

In this paper, we use a P-timed Petri Net to model a hypertext learning space and allocate a time delay pair (d^{min}, d^{max}) to each learning state or Petri Net place.

Some kinds of adaptations, like content, link, priority and timing, become possible by adjusting (d^{min}, d^{max}) conditionally. Some simple examples are given to illustrate the adaptation operations.

There are many situations in which this adaptation mechanism can be applied, for instance, temporal activity coordination in cooperative working environment. Also T-timed Petri Net (David, 1991) can be used to receive equivalent executive effect by linking time delay with transition.

Besides of timed Petri Net, colored Petri Net is an ideal tool to describe multiuser behavior in event-driven environment, where different users are granted different colored tokens. Transition enabling and firing is decided by classified colors held by place and color consuming function defined for each arc. In our case, if students at different levels are represented by different colored tokens, their access control becomes a matter of color allocation and token consumption.

In the future research, we are interested in applying colored timed Petri Net in the learning state space, also planning to introduce timed Petri Net based adaptation to collaborative environment.

References

IMC Information Multimedia Communication AG: Clix - learning management system. [Online]. Available: IMC AG: http://www.im-c.de/homepage/index.htm. (2005)

BLACKBOARD INC.: Blackboard course management system. [Online]. Available: Blackboard INC: http://www.blackboard.com/. (2005)

Brusilovsky, P., Eklund, J., and Schwarz, E.: Web-based Educations for All: A Tool for Development Adaptive Courseware. Proceedings of the Sevenths International World Wide Web Conference, WWW98. (1998)

Henze, N., and Nejdl, W.: Adaptation in open corpus hypermedia. IJAIED Special Issue on Adaptive and Intelligent Web-Based Systems vol.12. (2001)

Weber, G., Kuhl, H.-C., and Weibelzahl, S.: Developing adaptive internet based courses with the authoring system NetCoach. Proceedings of the Third Workshop on Adaptive Hypermedia, AH2001. (2001)

Peter Dolog, Nicola Henze, Wolfgang Nejdl, and Michael Sintek: Personalization in Distributed eLearning Environments. Proceedings of WWW2004 - The Thirteen International World Wide Web Conference, New York, USA, ACM. [Online]. Available: http://www.l3s.de/˜dolog/pub/www2004.pdf. (2004)

Dietrich Albert, et al.: Adaptive and Dynamic Hypertext Tutoring Systems Based on Knowledge Space Theory. Artificial Intelligence in Education: Knowledge and Media in Learning Systems, Volume 39 of Frontiers in Artificial Intelligence and Applications, Amsterdam, IOS Press. (1997)

Frank Halasz, et al.: The Dexter hypertext reference model. Communications of the ACM, 37(2). (1994) 30–39.

Rodrigo A. Botafogo, et al.: Identifying Aggregates in Hypertext Structures. ACM Proceedings of the Hyptertext'91 Conference, San Antonio, Texas, December. (1991)

Andrew Johnson, et al.: Adaptive Clustering of Scientific Data. Proceedings of the 13th IEEE International Phoenix Conference on Computers and Communication, Tempe, Arizona. (1994)

Shang Gao: A Web-based Hypertext Learning State Space Model and Learning Control. Proceedings of the Sixth IASTED International Conference: Internet and Multimedia Systems and Applications, Kaua'I, Hawaii, USA, August 12-14. (2002) 92–96.

P. David Stotts and Richard Furuta: Dynamic Adaptation of Hypertext Structure. Hypertext'91 Proceedings, Decemeber. (1991) 219–231.

Tadao Murata: Petri nets: Properties, analysis and applications. Proceedings of the IEEE, 77(4). (1989) 541–580.

Shang G., et al.: Supporting Adaptive Learning in Hypertext Environment: A High Level Timed Petri Net Based Approach. Proceedings of the ICALT 2005, to be appeared. (2005).

Exploring the Effective Search Context for the User in an Interactive and Adaptive Way

Supratip Ghose[1], Jason J. Jung[1], and Geun-Sik Jo[2]

[1] Intelligent E-Commerce Systems Laboratory,
School of Computer and Information Engineering, Inha University
Yonghyun-Dong, Nam-Gu, Incheon, 402-751, Korea
Sgresearch@gmail.com, j2jung@intelligent.pe.kr
[2] School of Computer and Information Engineering, Inha University
Yonghyun-Dong, Nam-Gu, Incheon, 402-751, Korea
gsjo@inha.ac.kr

Abstract. The explosive growth of information on the web demands effective intelligent search and filtering methods. Consequently, techniques have been developed that extract conceptual information from the document and use the conceptual information in the user profile to form part of the user's information intent from his/her query. In a similar vein, we build the profile without user interaction, automatically monitoring the user's browsing habits. These profiles, in turn, are used to automatically learn the semantic context of user's information need. These sets of categories can serve as a context to disambiguate the words in the user's query. In this paper, we present a framework for assisting the user in one of the most difficult information retrieval tasks, i.e., that of formulating an effective search query. Our experimental results show that implicit measurements of user interests, combined with the semantic knowledge embedded in a concept hierarchy, can be used effectively to infer the user context and to improve the results of information retrieval.

1 Introduction

The huge amount of data on the Web means that there is information on almost any topic available. Including personalization in Web information access may provide more effective information use and may help in retrieving the user information that matches their individual needs. Some systems for retrieving information try to include personalization in searching, browsing or both. In personalized search systems [1] [2], the search results are ranked according to user's interests or the searchable documents are arranged according to user-defined concepts for obtaining the desired information faster. An accurate representation of user's interests stored in some form of user profile is crucial to the performance of personalized search or browsing agents. The issues that need to be addressed in this process are how to build an accurate profile, particularly how to identify major or minor concept.

1.1 Background and Motivation

The word "ontology" is a specification of concepts and relations between them. It defines "content-specific agreements" on vocabulary usage, share, and reuse of

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 841–847, 2005.

knowledge. It is used to alleviate the communication problems between systems due to ambiguous usage of different terms. Usually, at the time of building user profiles, ontologies are used to address the so-called "cold-start problem". It means that systems happen to correspond poor performance at the time of initial learning of user interests and until those collect enough relevant information. Since initial behavior information is matched with existing concepts and relations between them, using ontologies as the basis of the profile helps to avoid or ease this problem. In our case, we use the Open Directory Project (ODP) [7] as our reference ontology. The "importance" of concepts and relations with respect to the search process can be explicitly stored to the ontology. User interaction to form queries in the ontology hierarchy may decide to form the "central" concept in which query reformulation can be benefited. Preferences for each individual user can be modeled by a search history. This can be used to fine-tune dynamically the list of concepts and relations displayed giving preferences to previously used ones.

2 Derivation of Concept Hierarchy from User Information Space

The search system in which we apply ontology relies on a sophisticated indexing process extracting taxonomy of related concepts from the raw documents. To do that, we take ODP concept hierarchy [and associated manually classified web pages] as our reference ontology. We choose the training documents from those associated with the concepts. The webpages, collected for a user are merged to create a collection of superdocuments sd_j and trained to form the index. The collection is obtained by preprocessing with stopword removal and stemming [4]. The vector-space model [4][6] is used to represent each term in the collections as vectors. Thus, each concept is treated as n-dimensional vectors in which n represents the number of unique terms in the vocabulary. Each term weight in the concept vectors is calculated using $tf \times idf$ and normalized by its length. In more detail, the weight of term i in concept j, is calculated as follows:

$$uw_{ij} = tf_{ij} * idf * cdf_{ij} \tag{1}$$

where sd_j = the superdocuemt used for training concept j, tf_{ij} = number of occurrences of t_i in sd_j and

$$idf_i = log \frac{\textbf{\textit{total no. of gathered concepts}}}{\# \textbf{\textit{ of concepts in the collection that contains term in }} t_i} \tag{2}$$

Where idf_i is the inverse concept frequency and

$$cdf_{ij} = log\left(\frac{\# of\ Training\ document\ for\ concept\ j}{\# of\ Training\ Document\ containing\ term\ i} \right) \tag{3}$$

where cdf_{ij} defines the concept document frequency, which is the number of training documents in the training set containing the terms [2].

The search system creates the user information space by collecting the documents during his browsing. This component works by passively observing the user's browsing behavior over time and collecting and analyzing documents in which the user has shown interests. The heuristics used by the system are based on several factors, including the frequency of visits to a page or a site; the amount of time spent on the page (or related pages within a site). In this phase, the URL's, time when visited, and webpages sizes are stored in log file by a proxy server. The program extracts the URL's for each user and filters them to remove those documents that are considered too short. Any webpage that represents user's information view is thus indexed to form the same vector space model. Consequently, the documents in the user's information space are classified by comparing the document vector to the representative vector for each concept. The top-similarity values, thus calculated, are then stored in the concept-based index.

2.1 System Architecture and User View Formulation

As shown in the Figure 1, the search system in our architecture uses a query module to allow the user to generate queries, an offline module to extract and learn the concept hierarchy representing user's view from the user information space, and an online module. The online module involves displaying the concept hierarchy to the user, allowing the user to interact over the indexed concept hierarchy as described in section 2.

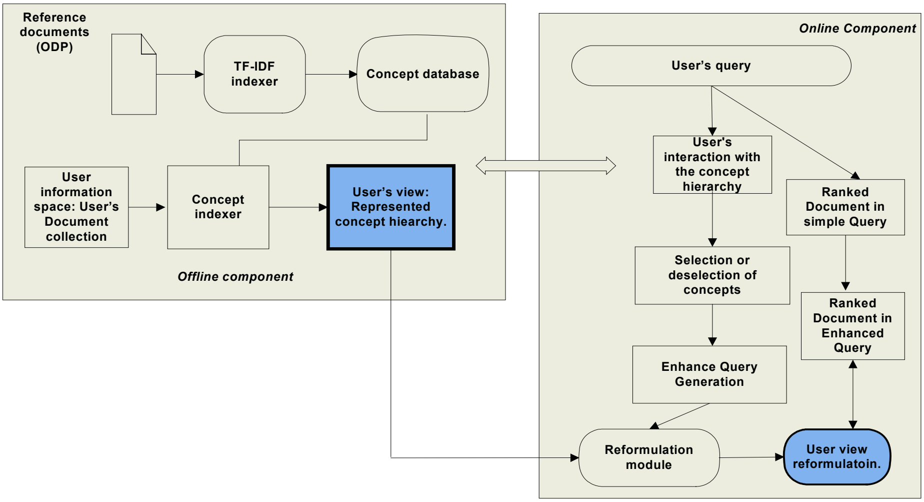

Reference documents (ODP)
TF-IDF indexer
Concept database
User information space: User's Document collection
Concept indexer
User's view: Represented concept hiearchy.
Offline component
Online Component
User's query
User's interaction with the concept hierarchy
Ranked Document in simple Query
Selection or deselection of concepts
Ranked Document in Enhanced Query
Enhance Query Generation
Reformulation module
User view reformulatoin.

Fig. 1. System Architecture

In order to initiate the query generation process, the user enters a keyword query. The query is represented as the same manner as in [6]. Based on the user's interaction with the system, the system responds by displaying the appropriate portions of the hierarchy. The system matches the term vectors representing each node in the concept hierarchy with the list of keywords typed by user. These nodes, which exceed a similarity threshold, are displayed to the user, along with other adjacent nodes. These nodes are the matching nodes of the concept hierarchy. Thereafter, user interface allows the user to provide explicit feedback.

The user expands those categories, which are relevant to the intended query, and deselect those categories, which are not relevant. We employ a variant of Rocchio's method [5] for relevance feedback to generate the enhanced query. The pre-computed term vectors associated with each node in the hierarchy are used to enhance the original query as follows:

$$Q_{en} = \alpha Q_{ort} + \beta \sum T_{sel} - \delta \sum T_{deselect} \tag{2}$$

In the above formula, T_{sel} is a term vector for one of the nodes selected by the user. On the other hand, $T_{deselect}$ is a term vector for one of the nodes, which is deselected by user. The factors α , β , δ are respectively the relative weight associated with the selected concepts, and the relative weight associated with the deselected concepts.

After the user gets the enhanced query, the user view is formulated with the formulation of the concept hierarchy. Once we got the weight related to the concept nodes then we can find the modified weight of the node by computing score of the node as follows:

$$\boldsymbol{Score(T_{i(n_i)}) = wgt_{rel} + \sum \mathbf{inf}(T_i, q, n_i)} \tag{3}$$

where w_{rel} is the weight associated to the node in consideration, n_i . $\inf(T_i, q, n_i)$ is the influence of term T_i for the selection of relevant nodes for the query. The measure $\boldsymbol{Score(T_{i(n_i)})}$ is defined here as the weight, that is formulated newly in user view to get the precise relevant documents.

3 Experiments with User View

The goal of this framework is to analyze the users' information need in the weighted concept hierarch in order to formulate the context of his/her search. Consequently, for the experiments we include top three levels of the ODP concept hierarchy. The user view is represented by the total weight and number of pages associated with each concept in the ontology. These sets of concepts are further restricted to include only those concepts that had sufficient training data (20 pages). The view was done on a daily basis for one month for total of six subjects and which three sets of users participated throughout the entire study. For this purpose, we created views consisting of level 1 concepts, views based on both levels 1 and 2, and views based on all three levels. For each view, we calculated the corresponding precision based on users' relevance judgments. Figure 2 shows the corresponding nature of precision. There accompanies a precision drop of 5% when the view is expanded from two levels to three. User nature of browsing can easily be viewed as users with less browsing habit find concepts in top two levels. Contrary to that, users with more browsing habits usually have many interests and several of level three concepts belonging to the same parent concept.

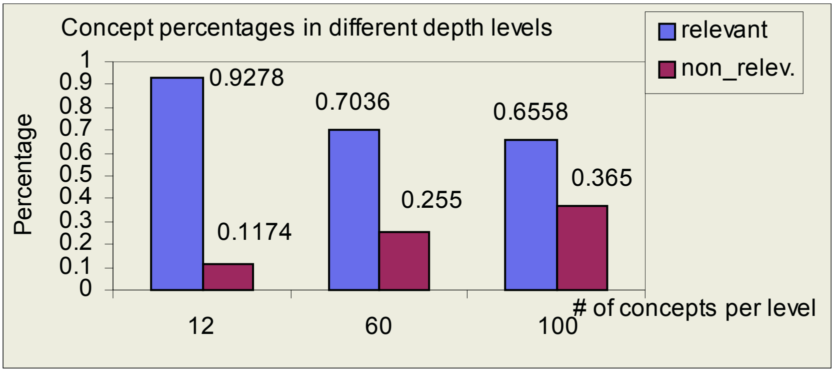


Fig. 2. Percentages of relevant and non-relevant concepts in different depth level

In order to evaluate our system, we compiled a list of keyword queries. The first set of keyword queries contained only one term such as fusion and costumes from the ODP ontology.

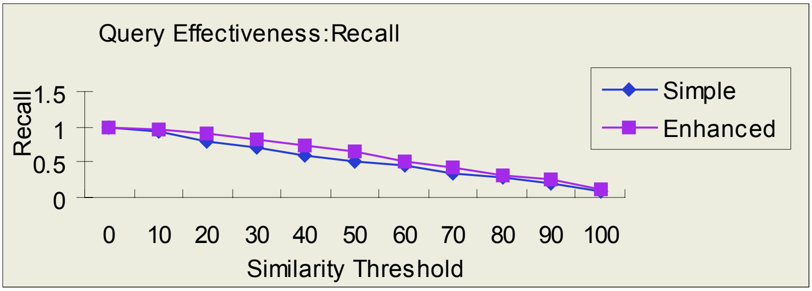


Fig. 3. Average recall for enhanced query on user view versus simple query search using two keywords

The second set of queries contained two terms such as cold fusion, cloths and costumes. We collected a number of "fitting" and a number of "noise" documents to construct our document collection. The "fitting" documents are those documents that should be ranked high in the search results. The "noise" documents are those documents that should be ranked low or excluded from the search results.

For example, if the intent for the search were to find documents about a Nuclear Fusion, the "fitting" documents would contain the fusion sense of the word fusion, while "noise" documents would contain the computers fusion, or Music related Jazz fusion as shown in Figure 5. If the user's ambiguous keyword causes the system to display several different portions of the hierarchy, the user interaction over the concept hierarchy modifies the domain view of the user. This will result in incremental generation of concept hierarchy over user's query. Like in Figure 5, it depicts the corresponding user is certainly interested in Nuclear fusion in atomic physics and thereby, should select the hierarchies related to that and deselect the node of Jazz fusion.

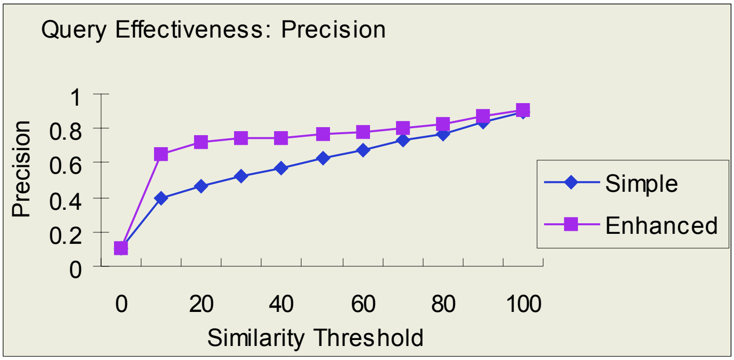


Fig. 4. Average precision for enhanced query on user view versus simple query search using two keywords

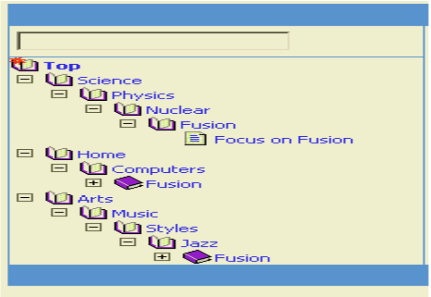


Fig. 5. Representation of the user view hierarchy interface

Using the expanded and deselected nodes, the system generates an enhanced query using the relevance feedback formula, which is described before. Based on the user's interaction with the concept hierarchy, the enhanced query that is generated accurately represents the user's intent for the initial query. We had the system to perform a simple query search and an enhanced query searches for each of our keyword queries. In the case of simple query search, a term vector was built using the original keyword(s) in the query text. In the case of enhanced query, we used the query that is generated after the user interaction over the concept hierarchy based on his/her relevance judgment depicted in Figure 2 and Figure 5. The search results were retrieved from the "fitting" and "noise" document collection by using a cosine similarity for matching. We have seen two types of improvements in the search results using the enhanced query. From the user's perspective, precision is improved since ambiguous query terms are disambiguated by the enhanced query. In addition, we have better recall in the search results since additional query terms retrieve documents that would not be retrieved by using only the original keyword query.

4 Conclusion

We have presented the implementation of query enhancement in this framework based on the user's interaction with a concept hierarchy. First by browsing, user view that represents user interest by concept. As the user's view is more improved over the concept hierarchy, we can let the user select the concepts to find the context of the query. We have shown the user preference of concepts in multiple level and we try to add those views in our future work to reformulate the query so as to minimize the gap between search and query formulation. Furthermore, a number of different statistical tests will be conducted to measure the strict validity of the standard performance measures on our framework.

References

1. Chen C., Chen M., Sun Y., PVA: A self adaptive Personal View Agent: Journal of intelligent Information Systems, p.173-194, 2002
2. Jason Chaffee, Susan Gauch. Personal Ontologies for Web Navigation. In Proceedings of the 9th International Conference on Information knowledge Management (CIKM), 2000, pp. 227-234
3. E. Glover, G. Flake, S Lawrence, W.Birmingham, A. kruger, C. Giles, and D. Pennock. Improving Category Specific Web Search by Learning Query Modifications. In Proceedings of the Symposiums on Applications and the Internet, SAINT 2001, San Diego, CA, January 2001
4. G. Salton and M.J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, NewYork, NY,, 1983
5. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, The SMART Retrieval System: Experiments in Automatic Document Processings, Pages 313-323. Prentice Hall, 1971
6. Yan, T. W. and H. Garcia-Molina, "Index structures for information filtering under the vector-space model," Proceedings of International Conference on Data Engineering, pp. 337-347, 1994
7. Open Directory Project. http://dmoz.org.

Generating CG Movies Based on a Cognitive Model of Shot Transition

Kazunori Okamoto[1], Yukiko I. Nakano[2], Masashi Okamoto[1],
Hung-Hsuan Huang[3], and Toyoaki Nishida[3]

[1] Graduate School of Information Science and Technology, the University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
{okamoto,kazu}@kc.t.u-tokyo.ac.jp
[2] Japan Science and Technology Agency (JST)
2-5-1 Atago Minato-ku, Tokyo 105-6218, Japan
yukiko@ristex.jst.go.jp
[3] Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
nishida@i.kyoto-u.ac.jp, huang@ii.ist.i.kyoto-u.ac.jp

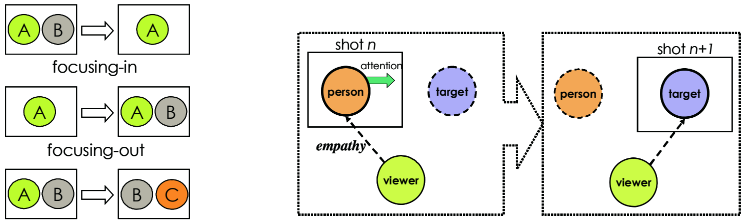
Abstract. This paper proposes a CG movie creation support system which automatically generates appropriate shot transitions to make the movies more comprehensible. First, we propose a shot transition model, and then, to examine the validity of the proposed model, we analyze shot transitions in TV programs. Finally, based on the data analysis, a CG contents creation system will be implemented.

1 Introduction

As computer graphics technologies have been greatly progressing, commercial CG software has been becoming popular and allowing non-professional users to create their own movies. Such CG tools are expected to support the users in expressing messages and ideas more visually and dynamically than text-based communication. However, videos shot by non-professional people are not well-organized and comprehensible. On the other hand, while filmmakers have cinematography techniques for making the scenes comprehensible for audiences, these techniques are usually learned only from experiences.

In order to generate such cinematographic expression automatically, a few previous CG graphics studies proposed knowledge-based approach to implement such implicit craft knowledge in a computer system [1, 2]. Instead of building a knowledge-base, in this paper we propose a method of generating movie contents featuring cinematographic expression based on a cognitive model of user involvement. Note that we are not concerned with artistic cinematography, but with enhancing comprehensibility in movie contents by establishing explicit models and rules of shot transition and implementing them in a computer system. Our system aims at supporting non-professional people to create movies comprehensible enough for the purpose of providing information.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 848–854, 2005.



(A) Cognitive Overlapping

(B) Empathy Channel

Fig. 1. Shot transition model

2 Background

As the basis of selecting appropriate shot transitions in producing comprehensible CG movies, we have proposed a theory of user involvement [3]. This section briefly describes this theory as the background.

2.1 Cognitive Overlapping

It has already been known that cognitively overlapped information is useful for understanding audio-visual signals as well as linguistic messages. For instance, our discourse and storytelling are based on a consistent flow of new and old information. For example, the following story clearly indicates such an information structure:

> *Once upon a time there was a king that wanted a new castle.* ***The king*** *hired the best castle builder in the land to build the castle. Prior to starting* ***to build the castle****,…*[4]

As shown in this example, linguistic overlapping enhances the comprehensibility of text. Moreover, in movies or TV dramas, it sometimes occurs that shots are changed while narration or actor's voice is still continuing on. Psychological research revealed that cutting (i.e., changing a shot) technique allows the audiences to find consistency between the shots and maintain the narrative structure across the shots despite visual discontinuities in the presented scene [5].

We call all these phenomena Cognitive Overlapping, and expect that the notion of Cognitive Overlapping would be useful for improving comprehensibility of CG movies. To apply this notion to shot transition selection in CG movies, we propose three types of shot transitions with cognitive overlapping (Fig.1-(A)), in which a part of the first shot remains in the second shot to maintain visual overlapping.

2.2 Empathy Channel

In non-overlapping transitions especially from a person shot to an object shot, a person in the previous shot frequently gives a pointing gesture, gaze, or verbal reference toward an object which lies out of the frame but is recognizable or accessible for her/him. Her/His attention behavior leads the audience to make mental contact to the hidden object through the attention as a reference point. As a result, the audience gets ready for the next shot featuring the object alone (see Fig.1-(B)).

In other words, through pointing gesture, gaze, or verbal reference by a person in a shot, the audience can empathize with the person and effortlessly relate the following shot to the previous one. We call such a communicative cue for attention functions the Empathy Channel.

3 Data Analysis

In this section, we examine whether shot transition models proposed in the previous section (i.e., the Cognitive Overlapping and the Empathy Channel) are actually observed in real movies. We analyzed three 30-minute TV programs. They provide information about popular items in a certain shop through a conversation between a TV host as guide and a shop owner as explainer.

We classified all the shots in the programs into the following seven shot types according to what was captured in each shot with respect to Cognitive Overlapping:

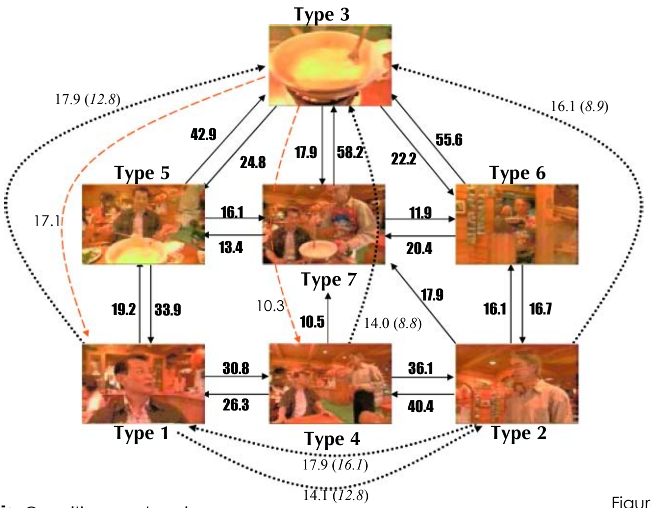
Shot Types

Type 1: The shot featuring the guide
Type 2: The shot featuring the explainer
Type 3: The shot featuring objects to be explained
Type 4: The shot featuring the guide and the explainer
Type 5: The shot featuring the guide and the objects
Type 6: The shot featuring the explainer and the objects
Type 7: The shot featuring the guide, the explainer and the objects

These programs consist of 485 shots, and 78 shots of them are categoried as Type 1, 56 shots as Type 2, 117 shots as Type 3, 57 shots as Type 4, 56 shots as Type 5, 54 shots as Type 6, and the other 67 shots as Type 7. As the programs we analyzed were those for the information about shops and their items, the most frequent shot type was Type 3, which features an object to be introduced to the viewing audience. Each of the transition rates from one shot to another is modeled in Fig. 2 (Note: the transitions between the same shot types and those below 10% occurrence are omitted).

This result shows that overlapping shot transition is frequently used in TV programs since transitions of this type occupies 77.9% of the whole transitions. Specifically, as illustrated in Fig. 2, *focusing-out* shot transitions are used to introduce new information to the audience, while *focusing-in* transitions to notify the audience of what s/he should pay attention to.

As for the rest of the cases, we found that 61.6% of the non-overlapping transitions are transitions from person shots to object/person shots, and 72.7% of those transitions include attention behaviors, such as a pointing gesture or a gaze toward the object(s) in the next shot. There results suggest that the concept of Empathy channel can account for non-overlapping transitions. Thus, Cognitive Overlapping and Empathy Channel cover 87.8% of the whole data.



gothic: Cognitive overlapping

(*italic*): Empathy channel (i.e. transition with gaze/pointing)

Figure: %

Threshold = 10%

Fig. 2. Shot transition network

4 System

Based on our theory of user involvement and the analysis of TV programs described in the previous sections, this section proposes a system that enables the user to create CG movie contents with little effort. The system consists of four main modules: the Content Editor, the Shot Generation Module, the Gesture Generation Module, and the Camerawork Generation Module. The details of each module are described as follows.

4.1 Content Editor

Through the Content Editor, the users specify scene information: (1) photos served as background of a CG content, and (2) utterances spoken by character agents. Users need to prepare only three photos: medium shot, left hand side background (BG1), and right hand side background (BG2). These three pictures construct a *virtual environment setting* (VES) in CG content as shown in Fig. 3. This is all what the users need to do in using this system. Movie contents are created automatically according to the information specified here.

4.2 Shot Generation Module

Then, in the Shot Generation Module, all the shot types defined in Section 3 are generated using the VES and three pictures chosen in the Content Editor. As shown in Fig. 3, the guide agent (G) and the explainer agent (E) are placed nearly face-to-face.

Shot type 1, 2, and 4-7 are produced by camera 1, 2, and 4-7 respectively. Shot type 3 is produced as a zoom-in shot from camera 7. Note that the two agents are joined with an *imaginary line* [6], which serves as a constraint on selecting possible camerawork.

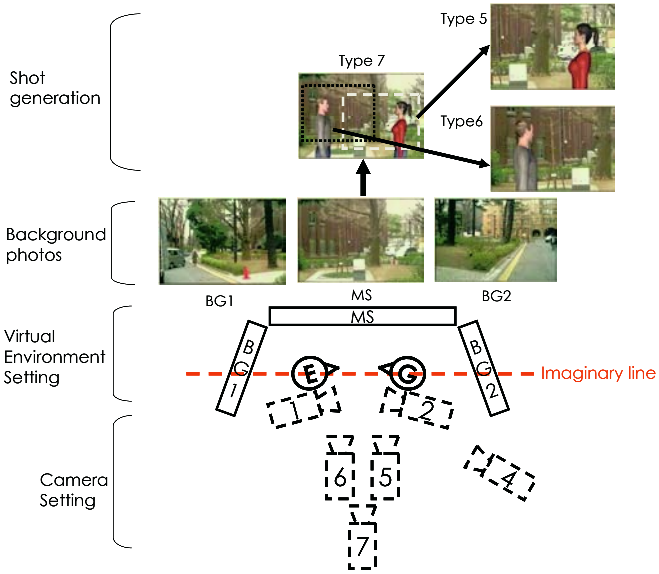


Fig. 3. Settings and shot generation from background photos

4.3 Gesture Generation Module

We employ the CAST system [7] as an agent Gesture Suggestion Module (GSGM), where the utterances specified in the Content Editor are analyzed and, according to the linguistic information in the utterances, candidates for agent behaviors (e.g., gestures and eye-gaze) are proposed.

When the Gesture Selection Module (GSLM) receives a behavior suggestion from GSGM, it selects appropriate gesture shapes according to the VES. At the same time, the GSLM receives gaze direction suggestions from Camerawork Generation Module based on the agent position in the VES.

Table 1. Shot selection rules

Next speaker	Referring to focused object	Shot type
Guide (G)	F	1, 4
Guide (G)	T	1, 3, 5, 7
Explainer (E)	F	2, 4
Explainer (E)	T	1, 3, 6, 7

4.4 Camerawork Generation Module

In the Camerawork Generation Module, camerawork is specified for each shot based on the shot transition model described in Section 2.

(1) Determining shot candidates: Shot candidates for each utterance are selected according to who is the next speaker and whether the utterance refers to the focused object or not. Rules for determining shot candidates are shown in Table 1. For example, if the speaker is the guide agent (G) and the utterance refers to a focused object (T), shot type 1, 3, 5, and 7 are selected as shot candidates.

(2) Generating shot transition: Among the shot type candidates chosen in step (1), this step selects one shot transition with highest probability in the shot transition network (Fig. 2). As major transitions produce Cognitive Overlapping, this step generates Cognitive Overlapping camerawork in most of the cases.

(3) Generating eye-gaze: When a selected transition lacks Cognitive Overlapping, Empathy Channel, which is the other device of user involvement, is produced using eye-gaze behavior by the speaker agent. The gaze direction is calculated according to a scene setting specified in the Content Editor.

5 Conclusion and Future Work

We have proposed a cognitive model of shot transition for CG movie contents, which consists of Cognitive Overlapping and Empathy Channel. Based on the model, we presented a system for CG contents creation, which supports non-professional users to create CG movies with little effort. As the result of the preliminary evaluation, we have gotten empirical support that the system can enhance the comprehensibility of the movies as long as they are based on our model.

However, we noticed that this system could be further improved in the following aspects: First, the users now select the three-picture sets that constitute of the background images of the movies via a typical file choosing dialogue box, this simple user interface lacks an overview toward the whole set of candidate materials and thus it is inconvenient and laborious for the users to determine the appropriate pictures in the content editor. Therefore, we are planning to develop a graphical front-end interface that provides an integrated environment of material storage, management and manipulation that associate actual spatial information of these material pictures. Users then can arrange their movie contents in a more intuitive way from position mapped background images.

Second, since the most obvious application of our system is to produce introduction videos of some places like sight-seeing spots, university campuses, etc., it is important to let the audiences easily capture the image of the spatial layout of those places from the videos. However, it is unclear whether the audiences can actually reconstruct the relative spatial positions of the objects appeared in their mind after watching the movies generated by the current system. This is because the current system focuses on processing transitions of camera positions within one scene rather than seamless connection of continuous scenes. We are interested in such issues and want to launch a further study in this aspect and improve our system in the future.

Third, the system and the description script language are in the early stage of development and are still relatively simple in its expressiveness and features, the scenes that a three background image set can express is limited, only two and just two characters are allowed, no other objects or properties can be included in the movie, there is no setting of the source of light thus unnatural shading can be seen in the generated movies, frequently used techniques of camerawork such as zooming and panning are absent and only three of them and seven camera positions are available, etc. We are going to solve these problems and extend our system to accommodate a larger variety of situations and scenes in the future.

At last, since our theory of user involvement remains a design theory that gives a speculative sketch for natural human computer interaction environment, it is desirable and expected to brush up the theory into that of evaluation.

References

1. Jinhong Shen, Seiya Miyazaki, Terumasa Aoki, Hiroshi Yasuda, Intelligent Digital Filmmaker DMP, *Proceedings of Fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'03)*, pp.272-277, 2003.
2. Doron Friedman, Yishai Feldman, Knowledge-Based Formalization of Cinematic Expression and Its Application to Animation, *Proceedings of Eurographics 2002*, pp.163-168, 2002.
3. Masashi Okamoto, Yukiko I. Nakano, and Toyoaki Nishida. Toward enhancing User Involvement via Empathy Channel in human-computer interface design. *Proceedings of IMTCI*, 2004.
4. http://www.umsl.edu/~sauter/analysis/fables/fall2002/king_castle.html
5. R. N. Kraft. Rules and strategies of visual narratives. *Perceptual and Motor Skills*. 64, 3-14, 1987.
6. J. May, M.P. Dean, and P.J. Barnard, Cinematography and interface design. In K. Nordby, P. Helmersen, D. J. Gilmore, and S. Arnesen, (Eds.), *Human-computer interaction: Interact 95*, 26-31. London: Chapman & Hall, 1995.
7. Y. Nakano, T. Murayama and T. Nishida. Multi-modal Story-based Communication: Integrating a Movie and a Conversational Agent, *IEICE Transactions, Special Issue on Human Communication (to appear),* 2004.

Analyzing Concerns of People Using Weblog Articles and Natural Phenomena

Toshihiro Murayama[1], Tomohiro Fukuhara[1], and Toyoaki Nishida[2]

[1] Research Institute of Science and Technology for Society
2-5-1 Atago, Mori Tower 18F,
Minato-ku, Tokyo, Japan
{tmuraya,fukuhara}@ristex.jst.go.jp
[2] Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, Japan
nishida@i.kyoto-u.ac.jp

Abstract. We describe an approach and some preliminary results of concern analysis using Weblog and natural phenomena data. By analyzing weblog article, we found relations between concerns of the people and weather data as natural phenomena factor.

1 Introduction

One of our motivations is to understand concerns of the people about social problem, because concern of the people is important in solving social problem. However, concerns of the people are various, changeable and moody, so understanding concerns of the people need analysis of huge and updated data. Figure 1 shows an overview of this paper. We analyze the weblog data for understanding relations between concerns of people and extra factors in the real world. Advantages of using weblog data are, (1)many weblog articles are written by personal-based, (2)weblog data provide time-series data (time-series weblog data are useful to analyze changes and periods of people concerns[1][2].). In this paper, we report our approach to find concerns via correlation analysis between time-series weblog data and factors which effect people. We consider that concerns of people are affected by various extra factors such as (1) natural phenomena represented by temperature and weather, (2) media information represented as news stories and TV programs, (3) human relations, (4) social situations, (5) cultural situations, and so on. In this paper, we focus on the relation between weather data as (1) natural phenomena.

This paper consists of following sections. In Section 2, we describe an overview of the system for collecting and analyzing Weblog articles. In Section3, we describe relations between time series weblog data and weather data. In Section 4, we discuss an application of the system in our society. In Section 5, we describe conclusion and future work.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 855–860, 2005.

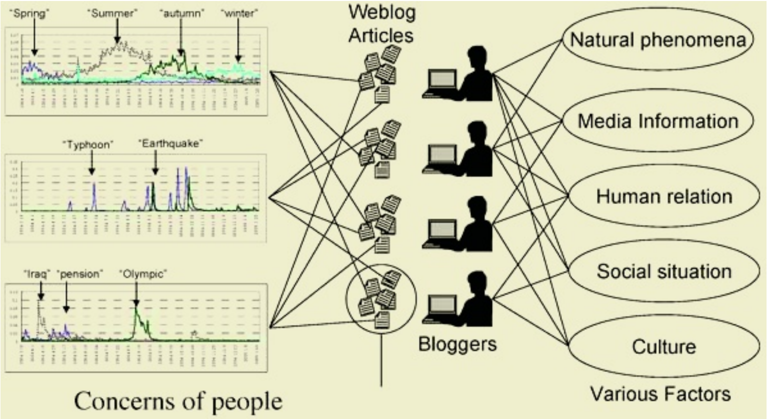


Fig. 1. Overview of this paper

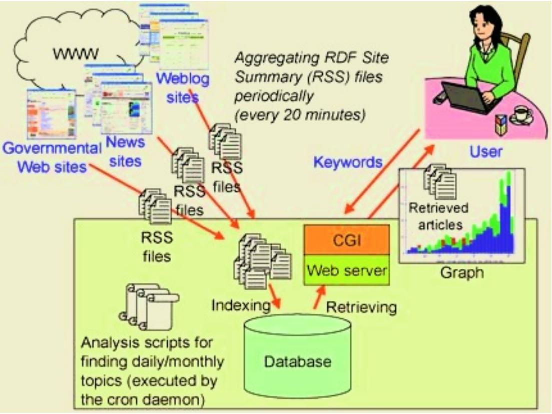

WWW
Weblog sites
News sites
Governmental Web sites
Aggregating RDF Site Summary (RSS) files periodically (every 20 minutes)
RSS files
RSS files
RSS files
Keywords
User
Retrieved articles
Graph
CGI
Web server
Indexing
Retrieving
Database
Analysis scripts for finding daily/monthly topics (executed by the cron daemon)

Fig. 2. System architecture of the KANSHIN

2 Weblog Data

For analyzing concerns of people, we created a system for collecting and analyzing Weblog articles called KANSHIN. The system collects Japanese and Chinese RSS files provided by Weblog sites. Figure 2 shows an overview of KANSHIN system. The system collects RSS files from Japanese and Chinese Weblog sites. An RSS file contains title, summary, date of publish, author, and category of an article. The system collects RSS files every 20 minutes. We collect about 20,000 articles per day for Japanese articles, and 1,000 per day for Chinese articles. We started to collect RSS files since March 18, 2004 for Japanese articles, and since January 25, 2005 for Chinese articles. By now, we collected 9,041,798 Japanese articles, and 41,671 Chinese articles. RSS files are collected from (1) personal Weblog sites, (2) news sites, and (3) governmental Web sites. In case of the first and the second type of sites, the system collects RSS files from Japanese Weblog ping servers such as Myblog Japan [3]

Table 1. Number of terms correlated with weather data

Weather data	Number of positive correlation term	Number of negative correlation term
average pressure	10	4
average sea-level pressure	12	7
average temperature	679	293
highest temperature	596	265
lowest temperature	721	284
average relative humidity	34	5
lowest relative humidity	31	3
average wind speed	2	0
maximum wind speed	14	0
maximum instantaneous wind speed	18	0
during daylight hours	3	15
rainfall level	88	0
hourly maximum rainfall level	74	0
ten minutes maximum rainfall level	78	0
amount of snowfall	103	0
maximum amount of snowfall	103	0
total	2566	876

directly. In case of the third type of information, the system acquires Web pages from governmental Web sites, converts them into RSS files, and acquires information from those files.

There are several related researches and tools such as blogPulse [4] and blogWatcher [5]. These services analyze Weblog articles, and provide histogram of words. One of differences is that our system analyzes concerns of people from collective and personal viewpoints [7]. Understanding both of social and personal concerns is important for understanding social problems.

3 Finding Relations Between Weblog Data and Natural Phenomena Data

There are various natural phenomenon factors such as temperature and weather, environmental index of air pollution, and so on. In this paper, we use time-series weather data of Tokyo in 2004 as a natural phenomenon factor. The period of data is 2004/03/18-2005/01/23. We calculated correlation coefficient between 16 weather data [6] (pressure, temperature, relative humidity, rainfall, snowfall ,wind speed, and so on) and term frequency data of Weblog articles.

In a result, Table 1 shows number of terms correlated with 16 weather data. There are two types of terms: (1) positive correlation terms, and (2) negative correlation

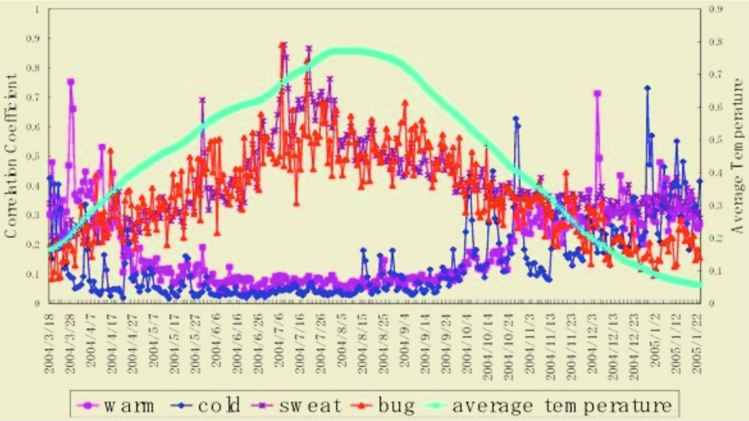


Fig. 3. Correlation between sample terms and average temperature

terms. For example, Figure 3 shows sample positive correlation terms (“bug”, ”sweat”) and negative correlation terms (“warm”, ”cold”) compared with average temperature. If temperature rises, frequency of positive correlation terms rise and negative correlation terms falls, if temperature falls, frequency of positive correlation terms falls and negative correlation terms rises. We found 2,566 terms which have positive correlation (correlation coefficient > 0.4), and 876 terms which have negative correlation (correlation coefficient < -0.4). This result means that natural phenomena such as temperature, rainfall, wind speed, snowfall and so on affect use of terms in weblog article. We found one another that some phenomena data such as temperature has both positive and negative correlation terms, others such as rainfall and wind speed has only one side. Table 2 shows a sample list of terms correlated with 16 weather data. Some of terms are within the scope of our prediction (e.g., “cool”, “warm”, “heavy snow” and so on), but some are not (e.g., “Bug”, “Mosquito”, “melancholy”, “leaking of rain” and so on). It is reasonable that number of insects increase according to the temperature, moreover high humidity may make people melancholy and heavy rain may cause leak in the roof. On the other hand, we couldn’t expect “Bug”, “Mosquito”, “melancholy” and “leaking of rain” until we see the result. Although we know that there are relations between weather data and terms such as “Mosquito”, we confirmed those relations objectively from this statistical result.

4 Discussion

In Section 3, we found relations between words appeared in weblog and natural phenomena data. For example, we found positive correlation between “water level” and “hourly maximum rainfall level”. This might mean that blogers paid attention to floods in case of abnormal rain. For another example, ”studless tire” and “maximum amount of snowfall” have positive correlation. This means that people have concerns of car accident caused by snow fall. We can find the point to which people paid attention by analyzing relations between terms and real world phenomena date. On the other hand, there are several points that most people are unaware of, but should be paid attention for avoiding and reducing natural disaster, such as insurance and preparedness.

Table 2. List of sample terms correlated with weather data

Weather data	Typical positive correlation term	Typical negative correlationterm
average pressure	“cold”, “runny nose”, “coat”	“sandal”, “cool”, “humidity”
average sea-level pressure	“cold”, “runny nose”, “coat”	“sandal”, “cool”, “humidity”
average temperature	“Bug”, “sweat”, “summer”	“warm”, “cold", “heater”
highest temperature	“pool”, “swimming” , “suntan”	“cold”, “Slope”, “Snowboard”
lowest temperature	“Mosquito”, “Stifling”, “Watermelon”	“Snowcapped mountain”, “Cold protection”, “Glove”
average relative humidity	“cloudburst”, “melancholy”, “cloudy sky”	“northerly-wind”, “warm”, “winter”
lowest relative humidity	“light rainfall”, “by bad luck”, “dull”	“warm”, “blowing snow”
average wind speed	“wind”, “big wind”	-
maximum wind speed	“big wind”, “voice of the wind”, “ocean waves”	-
maximum instantaneous wind speed	“shutter”, “south wind”, “the elements”	-
during daylight hours	“clear and sunny”, “sunlight”, “sunshine”	“cloudy sky”, “dull”, “light rainfall”
rainfall level	“rainfall", "soaking", "typhoon"	-
hourly maximum rainfall level	“rain coat” , “water level”, “leaking of rain”	-
ten minutes maximum rainfall level	“hectopascal”, “direct hit”, “cancellation”	-
amount of snowfall	“heavy snow”, “snow removal”, “snowy world”	-
maximum amount of snowfall	“gridlock”, “studless tire”, “soba”	-

5 Conclusion and Future Works

In this paper, we described our approach and preliminary results of concern analysis using Weblog articles. We described (1) a system for collecting and analyzing weblog articles called KANSHIN, (2) relations between weblog articles and extra factors such as natural phenomena. We found that natural phenomena affect use of terms in weblog articles. As future work, we have to investigate people’s attitude and awareness to natural disaster and social problems, and utilize this result effective information sharing between administration and peoples for reducing and preventing natural disaster.

References

1. Adler, S. The Slashdot effect. (online), 1999. (available from http://ssadler.phy.bnl.gov/adler/SDE/SlashDotEffect.html, accessed 2004-12-10).
2. Adar, E., and Zhang, L. Implicit structure and the dynamics of blogspace. In WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004. (available from http://www.blogpulse.com/papers/Adar_blogworkshop2_ppt.pdf, accessed 2005-03-03).
3. http://myblog.jp/ (in Japanese; accessed 2005-03-04)
4. Glance, N., Hurst, M., and Tomikiyo, T., BlogPulse: Automated trend discovery for Weblogs. In WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004. (available from http://www.blogpulse.com/www2004-workshop.html, accessed 2005-03-03).
5. Baeza-Yates, R. Modern information retrieval. Addison Wesley Longman Limited, U.K., 1999.
6. http://www.data.kishou.go.jp/index.htm (in Japanese; 2005-03-03)
7. Fukuhara,T., Murayama,T., and Nishida,T.: Analyzing Concerns of People using Weblog Articles and Real World Temporal Data, WWW2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation,Analysis and Dynamics, Chiba, Japan, May 10th (2005).

Sustainable Memory System Using Global and Conical Spaces

Hidekazu Kubota, Satoshi Nomura, Yasuyuki Sumi, and Toyoaki Nishida

Dept. of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan
{kubota,nomura}@ii.ist.i.kyoto-u.ac.jp
{sumi,nishida}@i.kyoto-u.ac.jp

Abstract. We present a concept and implementation for the computational support of spatial memory management and its temporal evolution. Our essential idea is using an immersible globe that consists of a global space and a conical space, which expands human biological memory. We developed Sustainable Knowledge Globe (SKG) for constructing a memory space, and then we have proposed a system called Contents Garden to expand SKG into immersive space. We have also proposed three perceptual operations on Contents Garden to improve operativity of SKG, and discussed them.

1 Introduction

The intelligent media technologies for capturing and utilizing our daily activities have been growing. We can leave traces of our activities by using a movie camera, a voice recorder and any other multimedia authoring tools; besides many studies make effort to capture our daily conversational situations by more sophisticated ways [1][2][3]. In a sea of captured records, sustainable memory system is indispensable for piling our knowledge upon knowledge. Here sustainability means development of a knowledge repository that accelerates the next development.

The purpose of this paper is to build a sustainable memory system that can co-evolve with human biological memory. Our essential idea is to develop a content space in a virtual landscape that expands human biological memory. The spatial world is generally a good container of a lot of objects. People can look over objects if they are arranged spatially. People can grasp a location of an object by using a lot of spatial clues such as left and right, high and low, near and far, and so on. Therefore the virtual landscape is expected to be effective for managing huge contents intuitively.

File management systems using spatial clues were discussed in [4][5][6], however their purpose is just a memory system. We more concentrate on the spiral development of memory, which is generally observed in knowledge processes like a conversational knowledge process [1] and SECI model [7].

We propose an immersible globe concept to facilitate development of content space. This concept looks like a memory forge that consists of a global space and a conical space. Both of them provide the space for building customizable intellectual world of a user. The global space gives a user an overlooking map for conceptual

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 861–867, 2005.

operations of contents. Meanwhile, the conical space gives an immersive field for perceptual operations. The two spaces can be switched by a user according to his needs. All contents on the global space can be projected to the conical space, and vice versa. The acceleration of the content development is achieved by repeating overlooking and immersive processes.

The immersible globe concept was firstly implemented on Sustainable Knowledge Globe (SKG) [8], and then we have proposed a system called Contents Garden to expand SKG into immersive space. This paper describes the concept and the systems for the computational support of spatial content building and its temporal evolution.

2 Immersible Globe Concept for Sustainable Memory System

We will now examine the immersible globe concept more closely. When we design and explore the space, we usually use two different styles: overlooking style and immersive style. Designing and exploring large buildings is a good example to illustrate our concept. A ground plan and an elevation are needed to grasp a structure of buildings. A perspective is also needed to imagine a view of the finished buildings. We can subjectively go through an immersive landscape using the perspective. In other words, we need a map and an experience of walking through to understand our town deeply.

The immersible globe concept also adopts two different styles that bring a user both objective and subjective views of the memory space. The user can layout contents anywhere on the space and investigates from contrasting views. We implement this concept by using a global space and a conical space correlated by Azimuthal Equidistant Projection (Fig 1). We adopt geographical metaphor to help user's spatial cognition. Azimuthal Equidistant Projection seems to be a good projection for a user not to get lost in both spaces because it keeps direction and distance.

The global space is an overlooking map like a terrestrial sphere. The user can browse whole contents from the objective (third person) view using conceptual interfaces such as typing keyboard, selecting menu and so on. Such styles are suitable for editorial works that needs overview and precise revision. On the other hand, the conical space is a ground that is immersive 3D world using perspective projection. The user can focus on specific contents in the subjective (first person) view. Immersive 3D world makes it easy to utilize perceptual user interface and talk with an embodied conversational agent (ECA) situated in specific content. ECA is expected to entrain participants to the content by conversational fashion.

These contrasts between the global space and the conical space are described on Table 1. Such overlooking-immersive approach would also be supported by zooming paradigm [9] in the visualization domain and SECI model [7] in the knowledge management domain. In SECI model, transformation of tacit/explicit knowledge facilitates knowledge creation. Here, subjectivity is one of the characteristics of tacit knowledge, and objectivity is that of explicit knowledge. Our approach is different from visualization and SECI model in that we focus to build long-term memory space in virtual 3D environment that expands human biological memory. The knowledge channel [10] is our former concept that facilitates knowledge evolution by using an editorial space and a conversational agent. This mainly focuses on conversational agent. The immersible globe reinforces the knowledge channel with expanding the memory space.

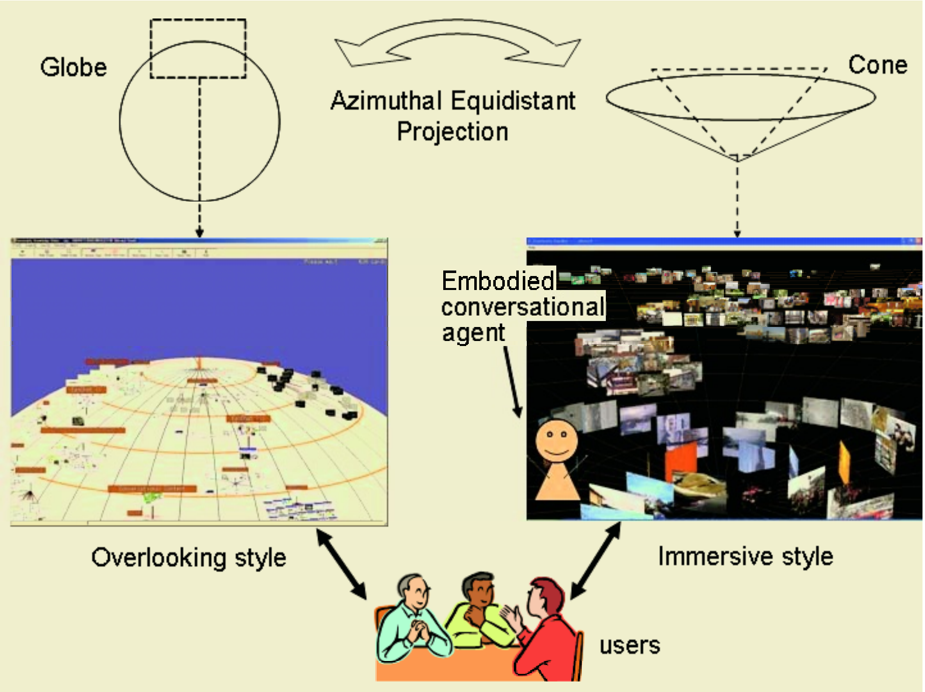

Globe
Azimuthal Equidistant Projection
Cone
Embodied conversational agent
Overlooking style
Immersive style
users

Fig. 1. Immersible Globe Concept

Table 1. Contrasts between the global space and the conical space

Shape / Attribute	Globe	Cone
Metaphor	Terrestrial sphere	Ground
Projection method	Parallel	Perspective
Distance from a user	Detached	Immersive
View	Objective	Subjective
Target	Whole	Focused
Interface	Conceptual	Perceptual
Interaction	Editorial	Conversational

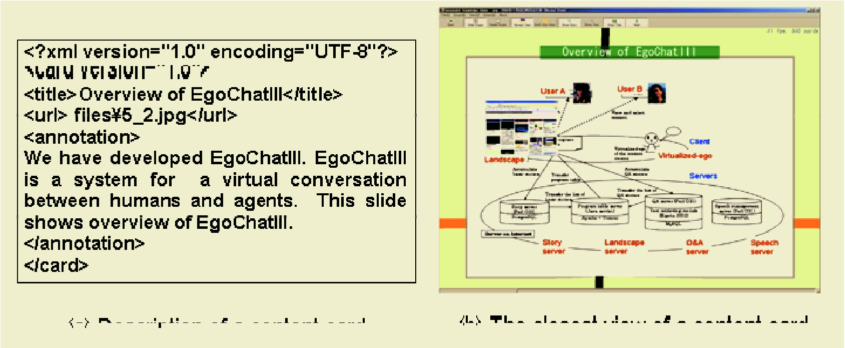
3 Sustainable Knowledge Globe (SKG)

SKG is sustainable memory system that has mainly objective view. SKG enables a user to edit contents on a global surface by using geographical arrangement, topological connection, and contextual relation. SKG also enables to look over huge contents by using a zooming interface [8].

SKG has a virtual globe like a terrestrial sphere that includes a sand-colored sphere with latitude and longitude lines, the landmarks for the north and south poles, and the equator (see the overlooking style in Fig. 1). There could be many kinds of topologies of the space: a finite plane, an infinite plane, 2-dimensional torus and so on. A finite plane seems to be difficult for a user to expand the content on the edges of the plane.

An infinite plane doesn't have this problem, however it is difficult for people to grasp infinite space. 2-dimensional torus also has no edge, however such topology may be unfamiliar for people. Therefore, SKG adopt a sphere. A spherical surface is a finite space, but people can arrange content on a sphere more freely than on a finite plane because a sphere has no edge. Moreover a sphere is more familiar shape than 2-dimensional torus because it likes a terrestrial globe.

Content on SKG is represented by a content card that consists of three parts: an URL of an embedded file, a title of a card, and an annotation of a card (Fig. 2 (a)). On the global surface, the content card stands upright with a thumbnail image (Fig. 2 (b)). The user can create cards anywhere on the globe, group the cards, and then operate them recursively by using tree structure. The user can also move around and zoom in/out the globe in search for interesting content. The rotation of the globe is restricted to along latitude and longitude lines because the non-restricted rotation of the globe often confuses people about the location and distance.



```
<?xml version="1.0" encoding="UTF-8"?>
<card version="1.0">
<title>Overview of EgoChatIII</title>
<url> files¥5_2.jpg</url>
<annotation>
We have developed EgoChatIII. EgoChatIII
is a system for  a virtual conversation
between humans and agents.  This slide
shows overview of EgoChatIII.
</annotation>
</card>
```

(a) Description of a content card

(b) The closed view of a content card

Fig. 2. An example of a content card

We have evaluated effectiveness of SKG through an experiment in personal situations [8]. Three subjects constructed their own memory spaces to manage their contents. The contents were mainly research slides and movies, and the rest of them are leisure photos, bookmarks and memos. The average number of content cards was 4,000. Here, various original arrangement patterns were observed on their memory spaces, such as a square, a star, a distorted world map, horizontal and vertical flows, a compact spiral and a clockwise arrangement. This experiment suggested that they reflected their own cognitive map onto these arrangements. It also suggested that the global space is not fit for focusing specific content, but suitable for overlooking contents.

4 Contents Garden

We have developed a novel system called "Contents Garden" that expands SKG into the immersive space. It aims to be able to focus the specific contents on SKG and operate perceptually in immersive environment. We equipped SKG with the global feature of the immersible globe concept, and then we have equipped Contents Garden with the conical feature. Whole contents on SKG can be projected onto Contents

Garden by Azimuthal Equidistant Projection (see the immersive style in Fig. 1). This projection centers a specific location on the globe and keeps the direction and distance of surrounding contents from the centered location. A terrestrial sphere generally projected onto a disk-shape, however we adopt a cone-shape not to conceal far contents by near contents. The inclination of the ground is 20 degrees. The user can arrange contents anywhere inside of the cone. The user can also explore the contents by moving and rotating the cone, zooming in/out the camera and changing the centered location by a mouse operation.

We propose three methods for operating contents perceptually. SKG is suitable for operating a strong connection like a tree structure, however not fit for operating ad hoc group of contents. We have got another suggestions from the experiment mentioned in the section 3. Below is a list of desirable ad hoc operations:

- - Making crowded contents sparse, and vice versa
- - Transforming broadly arranged contents into narrow arrangement
- - Making a space in crowded contents to add new contents
- - Moving a group of contents by pushing and pulling roughly

We have filled these demands by using three operations: “Distort”, “Make space” and “Push”. The user can select an operation from the context menu, and then indicate a target region by a mouse gesture that draws a closed circle.

Distort: Distort operation moves contents from a region to another region (Fig. 3 (a)). Each region is indicated by a mouse gesture. The contents in a source region are rearranged into a destination region keeping relative positions vertically and horizontally.

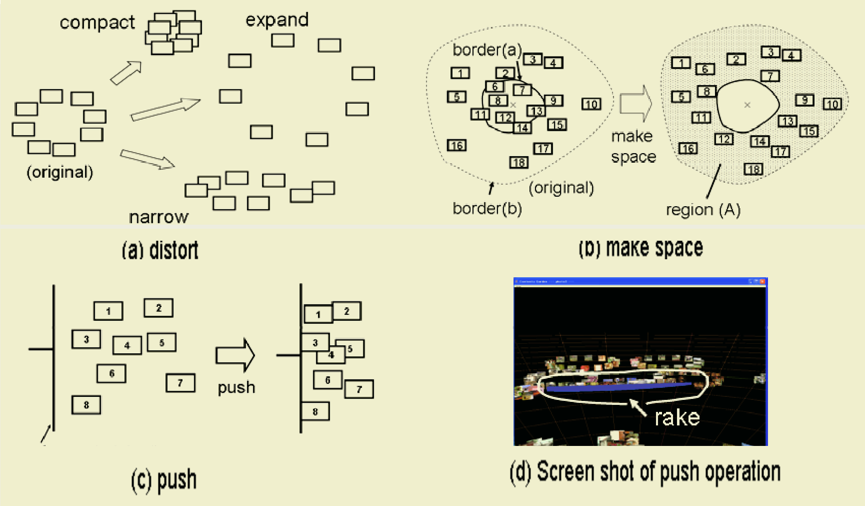
Make Space: Make space operation moves contents outward from a region (Fig. 3 (b)). This is an operation for making an empty area. The source region is indicated by a mouse gesture. The contents in source border (a) and extended border (b) (that is three times the size of border (a)) are rearranged into destination region (A) which is sandwiched between border (a) and border (b), keeping direction and near and far from the center of border (a).

Push: Push operation creates a broad rake to push and pull the contents (Fig. 3 (c)(d)). The size of the rake is indicated by a mouse gesture. The direction of the rake face can be rotated.

These perceptual operations are now in a developing state. We arranged 800 photos using these operations experimentally. “Distort” seems to be suitable for arranging jumbled contents because this operation doesn’t need a contents group structured in advance. The effect of “Make space” and “Push” are so unlimited that they easily destroy the surrounding arrangements. We plan to improve operativity of Contents Garden by using perceptual devices like a motion capturing system.

5 Discussion

We developed SKG for constructing a memory space, and then we have developed the Contents Garden for expanding SKG into the immersive space. They are individual systems, but they are data compatible. Now, we are planning to combine them for



(a) distort

(b) make space

(c) push

(d) Screen shot of push operation

Fig. 3. Three perceptual operations for positioning cards

a user to be able to build and explore contents switching the global space and the conical space smoothly. In the many related works about gestural interfaces for organizing spatially arranged contents [11] [12], Contents Garden focuses re-arrangement and management of huge contents, base on immersible globe concept. SKG and Contents Garden are now implemented on standard PC hardware. It is good for popular use, however an immersive environment with perceptual user interface is better for Content Garden to use its operations to advance. As yet our mouse gesture operations are simple, however they could be expanded by motion capturing system or another tracking system in the immersive environment.

6 Conclusion

We have presented the concept and the systems for the computational support of spatial content building and its temporal evolution. We developed Sustainable Knowledge Globe for constructing a memory space, and then we have developed Contents Garden for expanding SKG into the immersive space. We have also proposed three perceptual operations on Contents Garden to improve operativity of SKG, and discussed the advantage of “Distort” operation.

References

1. Toyoaki Nishida: Conversation Quantization for Conversational Knowledge Process, Special Invited Talk, S. Bhalla (Ed.): DNIS 2005, LNCS 3433, Springer, pp. 15-33, 2005.
2. Yasuyuki Sumi, Kenji Mase, Christof Muller, Shoichiro Iwasawa, Sadanori Ito, Masashi Takahashi, Ken Kumagai, Yusuke Otaka, Megumu Tsuchikawa, Yasuhiro Katagiri, and Toyoaki Nishida: Collage of video and sound for raising the awareness of situated conversations, International Workshop on Intelligent Media Technology for Communicative Intelligence (IMTCI 2004), pp.167-172, Warsaw, September, 2004.

3. Michihiko Minoh and Satoshi Nishiguchi: Environmental Media - In the Case of Lecture Archiving System, International Conference of Knowledge-Based Intelligent Information & Engineering Systems (KES2003), Vol.II, pp.1070-1076, 2003.
4. Joseph M. Ballay: Designing Workscape: an inter-disciplinary experience, SIGCHI conference on Human factors in computing systems: celebrating interdependence, pp. 10-15, 1994.
5. Stuart K. Card, George G. Robertson, and William York: The WebBook and the Web Forager: an information workspace for the World-Wide Web, SIGCHI conference on Human factors in computing systems: common ground, pp. 111-117, 1996.
6. George G. Robertson, Mary Czerwinski, Kevin Larson, Daniel C. Rob-bins, David Thiel and Maarten V. Dantzich: Data mountain: using spatial memory for document management, the 11th annual ACM symposium on User interface software and technology (UIST '98), pp. 153-162, 1998.
7. Ikujiro Nonaka and Hirotaka Takeuchi: The Knowledge-Creating Company, New York: Oxford University Press, 1995.
8. Hidekazu Kubota, Yasuyuki Sumi, and Toyoaki Nishida: Sustainable Knowledge Globe; A System for Supporting Content-oriented Conversation, in Proceedings of AISB 2005 Symposium Conversational Informatics for Supporting Social Intelligence & Interaction, pp.80-86, 2005.
9. Benjamin B. Bederson and James D. Hollan: Pad++: A Zooming Graphical Interface for Exploring Alternate Interface Physics, User Interface and Software Technology (UIST '94), pp.17-26, 1994.
10. Hidekazu Kubota, Jaewon Hur, and Toyoaki Nishida: Agent-based Content Management System. the 3rd Workshop on Social Intelligence Design (SID 2004), pp.77-84, 2004.
11. Thomas P. Moran, Patrick Chiu, William van Melle, and Gordon Kurtenbach: Implicit structure for pen-based systems within a freeform interaction paradigm, Proceedings of CHI'95, pp.487-494, 1995.
12. Elizabeth D. Mynatt, Takeo Igarashi, W. Keith Edwards, and Anthony LaMarca: "Flatland: New Dimensions in Office Whiteboards", ACM SIGCHI Conference on Human Factors in Computing Systems, ACM CHI'99, Pittsburgh, May 15-20, pp. 346-353, 1999.

Entrainment of Rate of Utterances in Speech Dialogs Between Users and an Auto Response System

Takanori Komatsu[1] and Koji Morikawa[2]

[1] Future University-Hakodate
116-2 Kamedanakano, Hakodate 041-8655, Japan
komatsu@fun.ac.jp
[2] Matsushita Electric Industrial Co, Ltd.
3-4 Hikaridai, Seika, Soraku, Kyoto 619-0238, Japan
morikawa.koji@jp.panasonic.com

Abstract. Entrainment, a physical phenomenon in which one individual's expressed information synchronizes with another's and vice versa, can be observed between two communicators who are in smooth interaction. In this study, we focused on the "rate of utterances" as communicators' expressed information, and then conducted an experiment to observe whether entrainments exist in the rate of utterances in speech dialogs between users and an auto response system. Specifically, participants were asked to read given dialog scripts with an auto response system that replied with different rates of utterances. The results revealed that 1) when the system's rates of utterances were faster, the participants produced faster rates of utterances, 2) when the system's rates were slower, participants spoke at slower rates. These results suggest the existence of entrainments of rates of utterances in speech dialogs between participants and an auto response system.

1 Introduction

Entrainment, a physical phenomenon in which one individual's expressed information synchronizes with the another's and vice versa, can be observed between two communicators who are in smooth interaction[1,2]. Recently, many researchers have developed various interactive robots designed to use entrainments to create smooth interactions with users [3,4].

For example, Ono et al.[5] focused on the role of the body (or gesture) entrainments in the communication between robots and users. Specifically, they conducted an experiment to observe how the entrained gestures of participants responded to the expressed gestures of a route direction robot and whether the participants could reach the destination. Their research revealed that most participants could reach the destination to which the robot guided them if their entrained gestures synchronized with those of the robot's. Watanabe et al.[6] developed a humanoid robot that can express nods and eyewinks in response to the user's speech. They reported that the users spoke in synchronization with the

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 868–874, 2005.

robot's nods and eyewinks. They concluded that the participants and this robot created smooth interactions by means of entrainments between users' speech and the robot's nods and eyewinks.

Heretofore, most studies that have focused on entrainments have investigated synchronization of body actions (e.g., gestures or nods) between users and artifacts; on the other hand, some studies have focused on entrainments in synchronization of speech sounds that users can easily exploit. For example, Nagaoka et al.[7] reported that the speakers in cooperative speech communications shared synchronized back-channel feedback or duration of utterances with their partners. Although this study was focused on the non-verbal information in speech dialogs, few studies have focused on the verbal information, such as speaking rate. In this study, we conducted an experiment to observe whether entrainments of "rates of utterances" exist in speech dialogs between users and artifacts. We focused on an auto response system (i.e., a telephone ticket reservation system) as an artifact with which users would interact.

2 Experiment

2.1 Participants

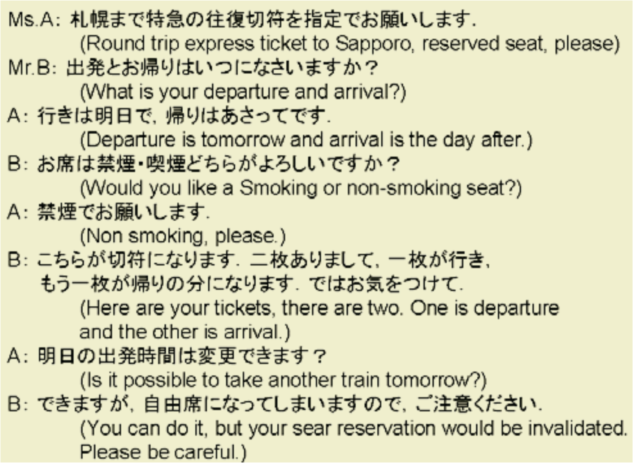
Participants were 27 Japanese (16 men and 11 women; 19–24 years old). Hearing test established that no participants had any hearing problems.

2.2 Setting and Procedure

First, an experimenter informed the participants that this experiment was to evaluate an auto response system and that the participants' task was to read given dialog scripts with this response system. We prepared three different scripts. The first script (Script Number 1) is about ordering train tickets (Ms. A is a traveler and Mr. B is a station officer, see Figure 1); the second script is about typical chatting between a high school girl and boy (like the dialog skits used in language schools), and the third is about ordering hamburgers at a fast-food shop (Ms. A is a customer and Mr. B is a shop clerk). The participants were asked to read Mr. B's part, while the response system played the speeches of Ms. A.

The auto response system used in this experiment just plays previously recorded speeches of Ms. A just 0.5 of a second after detecting the end of the human participants speech. This "0.5 second" was fixed so that the duration of turn-taking was constant throughout this experiment. As Ms. A's speeches, we used the recorded speech of only one female who read all three dialog scripts.

We used the sound authoring software "Cool Edit 2000" to prepare three different rates for Ms. A's utterances by expanding or contracting their durations without any modification of pitch values. Specifically, we prepared an 80% duration as the "High Speed" rate of utterance (H condition), 100% duration (no expansion or contraction of recorded sound) as the "Middle Speed" rate (M condition), and 120% duration as "Low Speed" rate (L condition).



Ms.A：札幌まで特急の往復切符を指定でお願いします．
(Round trip express ticket to Sapporo, reserved seat, please)
Mr.B：出発とお帰りはいつになさいますか？
(What is your departure and arrival?)
A：行きは明日で，帰りはあさってです．
(Departure is tomorrow and arrival is the day after.)
B：お席は禁煙・喫煙どちらがよろしいですか？
(Would you like a Smoking or non-smoking seat?)
A：禁煙でお願いします．
(Non smoking, please.)
B：こちらが切符になります．二枚ありまして，一枚が行き，
もう一枚が帰りの分になります．ではお気をつけて．
(Here are your tickets, there are two. One is departure
and the other is arrival.)
A：明日の出発時間は変更できます？
(Is it possible to take another train tomorrow?)
B：できますが，自由席になってしまいますので，ご注意ください．
(You can do it, but your sear reservation would be invalidated.
Please be careful.)

Fig. 1. Dialog script used in this experiment (Script Number 1)

Each participant experienced in nine dialogs (3 different scripts x 3 different rates for Ms. A's speech). We could then observe 243 dialogs (27 participants x 9 dialogs). In this paper, "the rate of utterances" was defined as the number of syllables in one speech divided by speech length [syllable/s]. These "rate of utterances" values were manually calculated from the recorded dialog sounds and dialog scripts.

2.3 Results

Analyzing 243 Dialogs. We divided the total 243 dialogs into three conditions (H, M and L) and calculated the participants' average rates of utterances in 81 dialogs for each condition (regardless of the kind of dialog script). The system's average rates of utterance were 11.533 [syllable/s] in H condition (81 dialogs), 8.83 [syllable/s] in M condition, and 7.74 [syllable/s] in L condition. While the human participants' average rate of utterance was 9.643 [syllable/s] in H condition; the average rate in M condition was 9.533 [syllable/s], and the rate in L condition was 9.188 [syllable/s] (Figure 2). Here, the results revealed that there were significant differences in these three conditions (F(2,160)=28.43, p<.01(**)). The Newman-Keuls test revealed a significant difference between H and L conditions (F(1,80)=45.67, p<.01 (**)) and between M and L conditions (F(1,80)=31.34, p<.01 (**)). This test also showed significant tendencies between H and M (F(1,80)=3.40, p<.0687(+)).

In sum, the participants adapted their rates of utterances to follow the system's rates of utterances so that this result suggests the existence of entrainment of the rate of utterances in speech dialogs between human participants and the auto response system.

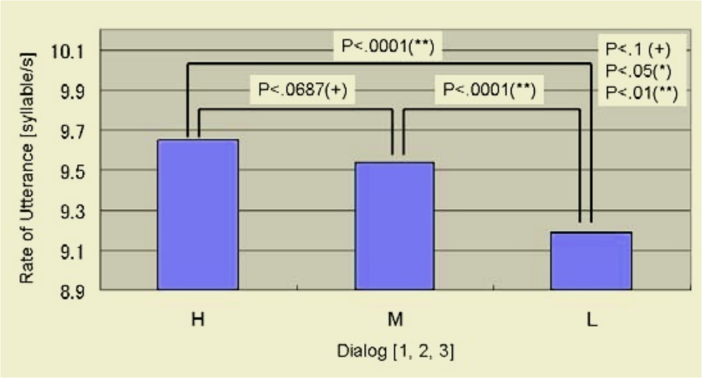


Fig. 2. Participants' rates of utterances in three different conditions (H, M, and L)

Analyzing 81 Dialogs in Script Number 1. As described above, we analyzed the 243 dialogs' average rate of utterances in each conditions regardless of the kinds of dialog scripts. Next, we observed whether the different scripts show different entrainment patterns.

At first, we will focus on Script Number 1 which is a dialog script about "buying a train ticket". The 81 dialogs performed using Script Number 1 were divided according to the three different conditions (H, M and L), and the average rates were calculated (Figure 3). The system's average rate of utterance were 11.356 [syllable/s] in H condition, 9.562 [syllable/s] in M condition, and 7.913 [syllable/s] in L condition.

The human participants' average rate of utterances when speaking with this system under the H condition was 9.562 [syllable/s], 9.441 [syllable/s] in M condition, and 9.011 [syllable/s] in L condition (Figure 3). This result revealed significant differences in speaking under these three conditions (F(2,52)=17.21, p<.01 (**)), and the Newman-Keuls test []indicated significant differences between H and L conditions (F(1,26)=45.67, p<.01 (**)), and between M and L (F(1,26)=16.59, p<.01 (**)). However, this test revealed no significant differences between H and M conditions (F(1,26)=1.83, n.s.).

Analyzing 81 Dialogs of Script Number 2. The 81 dialogs for Script Number 2, which is a dialog script of "everyday chatting between students," were divided according to the three conditions and the average rate of utterances for each was calculated. The system's average rate of utterance were 12.124 [syllable/s] in H condition, 9.942 [syllable/s] in M condition, and 7.989 [syllable/s] in L condition.

The participants' average rates were 10.096 [syllable/s] in H condition, 9.953 [syllable/s] in M condition, and 9.548 [syllable/s] in L condition. These results indicated significant differences among these three conditions (F(2,52)=12.45, p<.01 (**)). The Newman-Keuls test revealed significant differences between the H and L conditions (F(1,26)=21.95, p<.01 (**)) and between the M and L

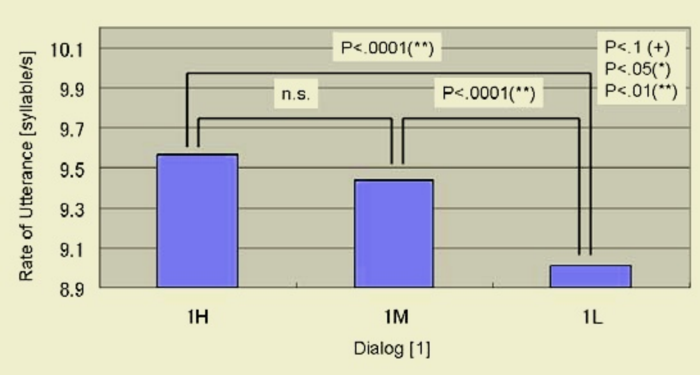


Fig. 3. Participants' rates of utterances for three different conditions of performing Script Number 1

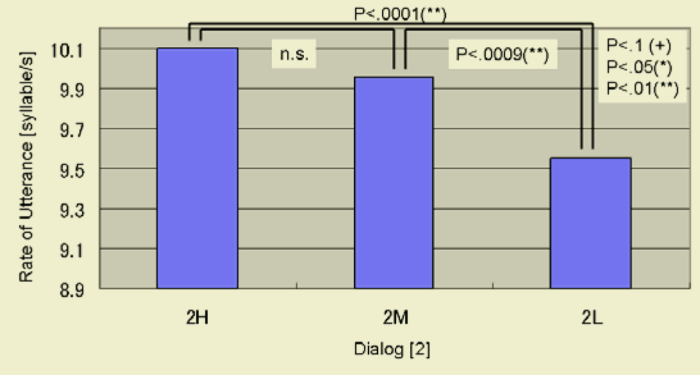


Fig. 4. Participants' rates of utterances in different three conditions of Script Number 2

(F(1,26)=14.38, p<.01 (**)), but revealed no significant differences between H and M conditions (F(1,26)=1.57, n.s.).

Analyzing 81 Dialogs in Script Number 3. The 81 dialogs in Script Number 3, which is a dialog script about "ordering hamburgers at a fast food shop," were divided according to the same three conditions and the average rate of utterances for each was calculated. Here, the system's average rate of utterance were 10.724 [syllable/s] in H condition, 9.001 [syllable/s] in M condition, and 7.060 [syllable/s] in L condition.

The participants' average rates were 9.272 [syllable/s] in H condition, 9.205 [syllable/s] in M condition, and 9.007 [syllable/s] in L condition. Similarly to Scripts 1 and 2, the results indicated significant differences among these three conditions (F(2,52)=3.48, p<.05 (*)). The Newman-Keuls test revealed that there were significant differences between H and L conditions (F(1,26)=4.80, p<.05 (*)) and between M and L (F(1,26)=4.38, p<.05 (*)), but revealed no significant differences between H and M conditions (F(1,26)=.25, n.s.).

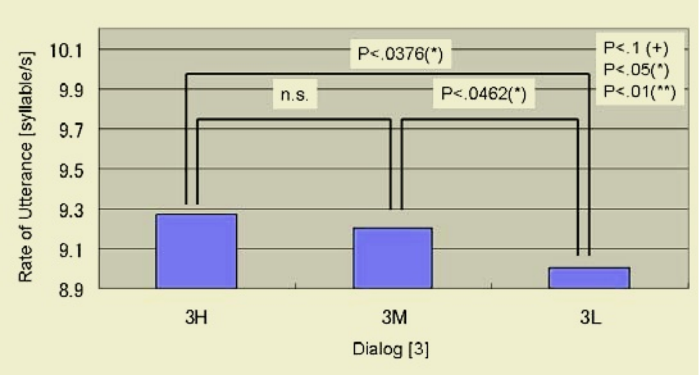


Fig. 5. Participants' rate of utterances in different three conditions of Script Number 3

3 Discussion and Conclusions

To sum up, the results of these experiments revealed that when the system's rates of utterance were faster (H condition), human participants produced higher rates of utterances, and when the system's rates were slower, humans used slower rates. This suggested the existence of entrainments of the rate of utterances in speech dialogs between human participants and the auto response system.

In the three different dialogs, although the significant differences in rate of utterances were found between 1) M and L conditions, and 2) H and L conditions, these differences were not found between H and M conditions. The reason for this last phenomena seemed to be that Ms. A's recorded utterances (used as M condition utterances) were originally faster; so that participants did not have sufficient "margin" to adapt their rate of utterance to the system's faster rates, but did have enough "margin" to adapt to the system's slower ones.

Figure 3, 4, and 5 showed that different dialogs have different average rates of utterances, i.e., the average rate of utterance in three conditions in Script Number 2 were generally fastest rates and in Script Number 3 produced lowest rates. The total of 243 dialogs were then divided into three scripts and the average rates of utterance for each script was statistically analyzed. We found significant differences among three dialogs (F(2,80)=35.40, p<.01 (**); Script 2 (fastest) > Script 1 > Script 3 (lowest)). The reasons for these phenomena are still unclear; however, we have at least three candidate explanations for this phenomena, and subsequent studies would be required to clarify these issues.

1. The recorded speeches (Ms. A's speeches) originally had different rates in each dialog, e.g., Ms. A spoke faster in Script 2 and slower in Script 3.
2. The effects of the "content" of each script were different, e.g., while the content of Script 2 facilitated the participants' faster rates of speech, the content of Script 3 did so for the lower rates.
3. Some possibilities are that this study's definition of the rates of utterances cannot handle the participants' bloopers. To resolve this issue might require introducing the number of morae instead of the number of syllables to calculate the rates of utterance.

In addition, we are planning a subsequent study to observe how participants can adapt their rates of utterance to the system's rates that change drastically, e.g., the system plays the H condition recorded sound, and then in turn plays the L condition sound. If participants could change their rates of utterances to follow the system's drastically changing rates at every turn, we could strongly argue the existence of entrainment of the rate of utterance in speech dialogs between users and an auto response system.

References

1. Condon, S. W. and Sander, L. W.: Neonate movement synchronized with adult speech: Interaction participation and language acquisition; Science, Vol. 183, 99–101 (1974).
2. Watanabe, T. and Okubo, M.: Evaluation of the Entrainment Between a Speaker's Burst-Pause of Speech and Respiration and a Listener's Respiration in Face-to-Face Communication; In Proceedings of RO-MAN'97, pp. 392–397 (1997).
3. Wesugi, S., Katayama, T., and Miwa, Y.: Virtually Shared "Lazy Susan" Based on Dual Embodied Interaction Design; In Proceedings of INTERACTION 2004, pp. 263–270 (2004).
4. Miyake, Y.: Co-creation system; Cognitive Processing, vol. 3, pp. 131–136 (2002).
5. Ono, T., Imai, M., and Ishiguro, H.: A Model of Embodied Communications with Gestures between Humans and Robots; In Proceedings of the 23rd Annual Meeting of the Cognitive Science Society (CogSci2001), pp. 732–737 (2001).
6. Watanabe, T., Okubo, M., Nakashige, M., and Danbara, R.: InterActor: Speech-Driven Embodied Interactive Actor; International Journal of Human-Computer Interaction, Vol. 17(1), pp. 43–60 (2004).
7. Nagaoka, C., Komori, M., Draguna, M., Kawase, S., Yuki, M., Kataoka, T., and Nakamura, T.: Mutual Congruence of Vocal Behavior in Cooperative Dialogues: Comparison between Receptive and Assertive Dialogues (In Japanese); In Proceedings of Human Interface Symposium 2003, pp. 167–170 (2003).

Locomotion Control Technique for Immersive Conversation Environment

Rai Chan, Jun Takazawa, and Junichi Hoshino

Systems & Information Engineering , University of Tsukuba Tsukuba Ibaraki , Japan
jhoshino@esys.tsukuba.ac.jp

Abstract. Generating composite human motion such as locomotion and gesture is important for interactive applications, such as interactive storytelling and computer games. In interactive story environments, characters do not merely stand in one position, they should be able to compose gestures and locomotion based on the discourse of the story and other object locations in the scene. Thus in this paper, we propose a conversational locomotion model for virtual characters. We constructed a conversational locomotion network for a virtual environment. A multi-pass searching algorithm calculates the optimal walking path, which uses node activation from the story locations and conversation units. The character also locally adjusts its position so that it does not limit the referenced object from the user's sight. We have applied our technique to the interactive 3D movie system, and demonstrated the composite motion of the character's locomotion and conversation thus strengthens the immersion in the story environment.

1 Introduction

In our daily life, humans perform many composite actions simultaneously. Walking and talking concurrently is one of the typical composite human actions. Composing locomotion and gestures is also important for applications such as interactive movies and games. In the interactive story environment, characters do not simply stand in one position, they should be able to compose gestures and locomotion based on the series of story locations and surrounding objects.

The proper location and timing of the character is influenced by various contexts, such as the connection of scene locations and the current environment. The apparent size of objects and the detail of the explanation affect how much closer the character should move. Connection of the scene locations also affects the current position. When the character refers to particular objects during a conversation, and the objects are far from the character's current position, it is time consuming to make the character approach the object every time. However, when the referenced object is close to the next scene location, it is more reasonable that the character moves closer to that object.

In this paper, we propose a mechanism for fluid conversational locomotion for virtual characters. This is realized by calculating the optimal locomotion path, which is influenced by the conversation and the story location, and the characters then subsequently generate composite walking and conversation actions. The character also

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 875–882, 2005.

locally adjusts its position so as not to limit the referenced object from the user's sight. Figure 1 shows a typical example of conversational locomotion. In this scenario, the character first explains that the vending machine cannot be approached in the disaster zone. In the next scene, the user asks about the other specific objects that cannot be approached, and the character then moves closer to explain more about the background relating to the disaster.

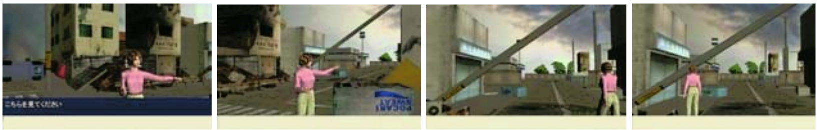


Fig. 1. Example of conversational locomotion. The character generates composite locomotion and gestures using story locations and local conversations

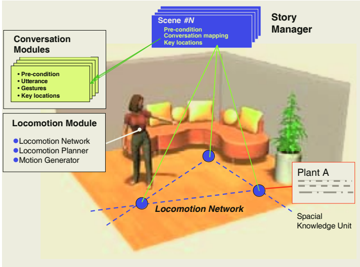
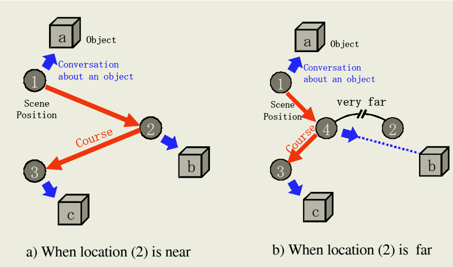


Fig. 2. Overview of the conversational locomotion architecture



a) When location (2) is near

b) When location (2) is far

Fig. 3. A concept of conversational locomotion using a simple example

2 Synchronization of Conversation and Locomotion

2.1 Overview of the Architecture

Figure 2 shows the conversational locomotion architecture. The system has a locomotion module, conversation modules, and a story manager. A story consists of a set of scene units and this controls the discourse of the conversation. A scene unit has a precondition, scene location, and links to a collection of possible conversation modules. Proper scene units are selected using the preconditions, such as a change in environment and the history of the user's verbal expressions. When a story unit is selected, the possible conversation units applied in the scenes are activated.

Conversation modules have preconditions, utterance, corresponding gestures and key locations. The locomotion module dynamically plans locomotion paths and generates walking motion patterns based on key locations.

2.2 Key Location Control

To compose locomotion and conversation, we need to decide the character's location and the timing of walking during the conversation. The locations of the actors are influenced by where the scenes are taking place and the content of the conversations. Figure 3 shows a typical example of locomotion planning during a conversation. Assuming that the actor should move from node1 to node3, it is reasonable that the actor stops at node4 if the referenced object is visible enough.

Locomotion and conversation are composed by considering the following three types of location constraints:

1) Scene location: The scene location corresponds to where the actions and conversations are taking place. To begin a conversational scene, the actor should be at a proper location.
2) Interpersonal location: The character changes relative locations from the other actors during conversation. For example, when the character begins to talk it needs to approach the other participants. When the character tries to explain something, visibility of the referenced object is also considered to decide interpersonal location.
3) Reference location: This is the relative location of the character and the referenced object.

These location constraints are used as key locations k_n in the conversational locomotion planning. The key location consists of a position in the floor coordinate system, and a standing duration t_k at a given key location. In most scenes, the proper standing position of the character has a degree of freedom. A key location has a several candidate positions with different activation values.

The standing duration of the key location can be dynamically changed by the key location control rules in the conversation units. For example, the initial standing duration can be used to decide how long the character can talk with the user at that particular position. When conversation with the user ends, the conversation units set the standing duration to zero that then causes the character to move onto the next scene location.

2.3 Conversational Locomotion Network

Activating the locomotion network using story locations and conversation units controls conversational locomotion. An optimal locomotion path is selected by calculating the optimal locomotion path with the maximum activation.

The locomotion node k represents a point on the floor coordinate systems (u_k, v_k). Characters can walk away from locomotion nodes for local position adjustment. The locomotion network $N_k = (G_k, length_k)$ consists of directed graph $G_k = (K, E)$, where the edges representing distance between the nodes are represented as $length_k(e)$. The initial locomotion network is constructed by sampling the possible standing locations. The candidate node positions are story locations and objects that have been referenced previously in conversation units. To increase the possible locations, we randomly sample the possible walking space.

Associating a key location at a proper clause in utterance controls timing of locomotion. For example, the reference location can be associated with clauses including the referenced object. There are several methods for associating the key location to a clause. When the numbers of conversational modules are limited manual association may be relatively easy. Even if the key location specification is predetermined, the actual character motion is dynamically changed depending on the story locations and the order of the conversations.

Timing of utterance is also synchronized to character locomotion. As illustrated in Figure 3b, the pre-condition of conversation units are used to pause and thereby wait until the character moves to the proper positions.

3 Conversational Locomotion Planning

Key locations are activated using scene locations and activation rules in the conversation modules. The locomotion path is dynamically selected by using a multi-pass searching algorithm that calculates the maximum activated path. When the conversation units change the status of activation the locomotion path is recalculated.

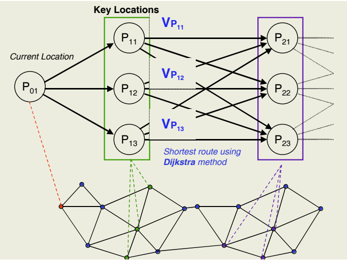


Fig. 4. Key location and multiple pass searching. N-best key location is selected to search the maximum activated pass

3.1 Multiple Pass Searching

Multiple key locations are set with a different activation value. By selecting N-best key locations, the possible locomotion segments between key locations are selected. The total activation along the locomotion segments is calculated. Locomotion segments between candidate key locations are obtained by employing the method of Dijkstra described previously [Dijkstra 1959]. As shown in Figure 4, we calculated candidate locomotion segments such as P_{00}- P_{01} and P_{00}- P_{02} ,to obtain the total activation value.

3.2 Activation Functions

In addition to the scene locations, we use the apparent object size and walking size to locally control locomotion. Figure 5 (a) and (b) shows the activation function used in this system.

1) Apparent object size: $A(P_{s,n})$

When the apparent size of the referenced object is small, the character should move closer until it becomes large enough to visualize easily. We determined the activation function as shown in Figure 5a. A sphere approximates the reference object, and the view angle from the user's eye position is calculated. Note that the approximated object size corresponds to the object area referred to in the conversation. When the character refers to a small area of a big object, the approximated object size is small. Orientation constraints are also integrated by forming activation distribution to a specific direction.

2) Walking distance: $D(P_{s,n}, P_{s+1,m})$

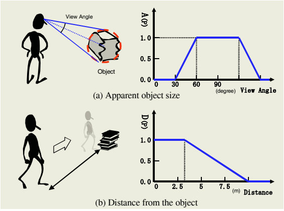
When the walking distance from the current location of the character is longer, the character tries to avoid this longer path. We determined the activation function as depicted in Figure 5b.

The total activation values are calculated along locomotion segments.

$$V(P_{0,0}, P_{1,n_1}, \ldots., P_{s,n_s}) = \sum_{t=1}^{s} \left\{ w(t) \cdot \left[\alpha A(P_{t,n_t}) + (1-\alpha) D(P_{t-1,n_{t-1}}, P_{t,n_t}) \right] \right\}$$

Where w(t) is a weighting value. w(t) is used to control the number of key locations that the character should consider.

Another type of activation function is easily integrated into this framework. For example, the access control of the character to a specific area can be represented. By setting the negative activation value to the specific locations, the character will avoid entering that place.



(a) Apparent object size

(b) Distance from the object

Fig. 5. Scene constraints and activation value. The apparent object size and distance form the object causes trade-off

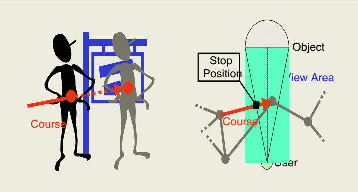


Fig. 6. The local position adjustment using the user's relative position and the referenced object

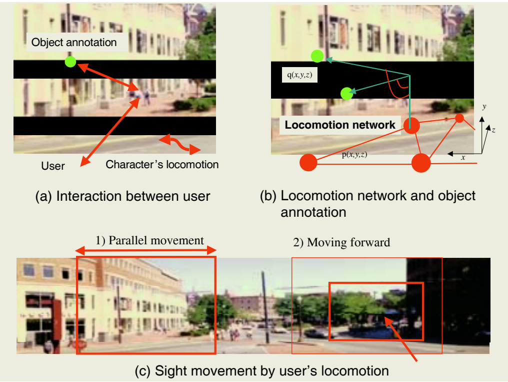
3.3 Local Position Adjustment

The actor's position is locally adjusted so as not to obscure the user's sight of the referenced object. Figure 6 shows the concept of local position adjustment. As described in section 3.2, we approximate the reference area by using a sphere. The viewing area of the user is calculated from the 3D location of the user's eye and the reference object sphere. When the character approaches the object, it stops at the intersection of the view area and the edges of the locomotion network.

4 Panorama-Based Immersive Story Environment

4.1 Introduction

As it is well known that constructing convincing photo-real 3D models is very time consuming, we first attempted to build a panorama-based immersive story environment to prove our theory. Virtual characters can walk and talk in a photo realistic environment by using a combination of locomotion network and object annotations. A model of the environment is approximated by the linked panorama images. The character and the user can thus move around in the photo-real scene.



(a) Interaction between user

(b) Locomotion network and object annotation

(c) Sight movement by user's locomotion

Fig. 7. Interaction mechanism for panorama-based conversation environment

4.2 The Generation of Node Coordinates in 3D Space

To facilitate the character walking in the environment, we need a correspondence between the 2D panorama image and its 3D environment. We constructed the locomotion network to determine the walking area and calculate the appropriate walking path. Figure 7b shows an example of the locomotion network. Three-dimensional locations of scene objects are annotated so that the character can refer to these and point during conversations.

When the camera coordinates $\mathbf{P}_c = (x_c, y_c, z_c)$ and the 2D coordinates of the node in the panorama image are assumed to be $\mathbf{P}_n^{2D} = (x_n^{2D}, y_n^{2D})$, the 3D space where a virtual character exists $\mathbf{P}_p^{3D} = (x_p^{3D}, y_p^{3D}, z_p^{3D})$ is converted into the projection coordinates by

$$x_p^{3D} = x_c^{3D} \tag{1}$$

$$y_p^{3D} = \frac{2C_y}{H} y_n^{2D} \tag{2}$$

$$z_p^{3D} = f \tag{3}$$

When a virtual character was displayed in the panorama image, it was necessary to convert the projection coordinates $\mathbf{P}_p^{3D}$ into the coordinates $\mathbf{P}_n^{3D}$ of 3D space to present the image that moves in the direction as depth changes.

The straight-line equation extracted from the camera coordinates $\mathbf{P}_c$ to projection coordinates $\mathbf{P}_p^{3D}$ is represented by:

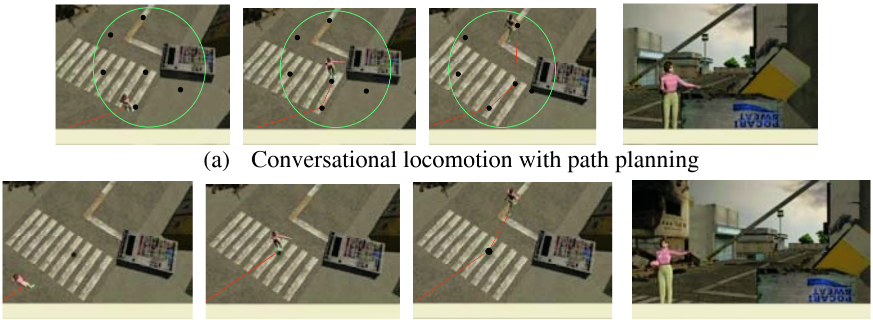
$$y = \frac{y_p^{3D}}{f}(z - z_c) + y_c \tag{4}$$

By employing the expression (4), when either the z or y value is determined, the other value is also determined. Then, the z value is set as a value proportional to height in the panorama coordinates by the use of the next expression.

$$z_n^{3D} = \left(z_{\max}^{3D} - z_{\min}^{3D}\right)\frac{\left(y_n^{2D} + \frac{H}{2}\right)}{y_{\max}^{2D}} + z_{\min}^{3D} \tag{5}$$

5 Results

Furthermore, we have applied our composite motion generation technique to interactive 3D movie applications. As mentioned earlier, in the interactive story environment actors do not simply stand in one position, as they constantly interact with their environment. If the actor changes its responses by the user's input, the actor should then decide when and where to move.



(a) Conversational locomotion with path planning

(b) Conversational locomotion without path planning

Fig. 8. Snapshots from a conversational locomotion sequence, illustrating the concept of locomotion planning

5.1 Composition of Locomotion and Conversation

Figure 8 shows the result of conversational locomotion. To synchronize locomotion and conversation, we need to determine location and timing of locomotion during conversation. Figure 8a shows a snapshot with locomotion planning, while Figure 8b illustrates a snapshot without locomotion planning. We can see that the actors try to select a closer position to explain relevant details.

5.2 Locomotion Conversation Based on Panorama Picture

A panorama of images surrounding a church (Figure 9) was constructed to confirm the effectiveness of the technique. In this panorama image, the technique mentioned previously was applied, and a virtual character moved in the panorama space while simultaneously talking with the user. Thus a movie where a virtual character acted as a guide of the building was generated very effectively (Figure10).

Fig. 9. Panorama image of church

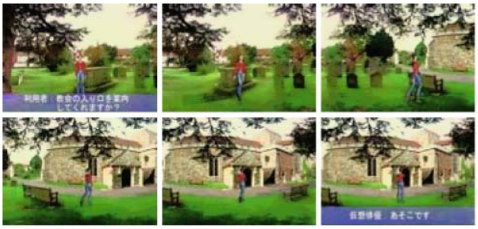


Fig. 10. Snapshots from virtual trip sequences

6 Conclusion

In this paper, we have proposed a conversational locomotion model for virtual characters. By calculating the optimal locomotion path influenced by conversation and story locations, the characters generate composite walking and conversation actions. The character also locally adjusts the position considering the visibility of relevant objects from the user. We have validated our model, and produced a highly accurate interactive animation sequence using conversational locomotion that will provide significant improvements to current interactive story applications.

References

Dijkstra, E. W. 1959. A note on two problems in connection with graphs. NumerischeMathematik1, 269–271.

Bandi, S. and Thalmann, D. 1998. Space discretization for e_cient human navigation. *ComputerGraphicsForum17,*3, 195–206.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M., 1994. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In Proceedings of ACM SIGGRAPH '94, 413–420

Presentation of Human Action Information via Avatar: From the Viewpoint of Avatar-Based Communication

Daisaku Arita and Rin-ichiro Taniguchi

Department of Intelligent Systems, Kyushu University
6-1, Kasuga-koen, Kasuga, Fukuoka, 816-8580, Japan
{arita,rin}@limu.is.kyushu-u.ac.jp

Abstract. This paper describes techniques to present human action information on an avatar-based interaction system, using real-time motion sensing and human action symbolization. Avatar-based interaction systems with computer-generated virtual environments have difficulties in acquiring user's information, i.e., enough information to represent the user as if he/she were in the environment. This mainly comes of high degrees of freedom of human body and causes the lack of reality. Since it is almost impossible to acquire all the detailed information of human actions or activities, we, instead, recognize, or estimate, what kind of actions have occurred from sensed human motion information and other available information and re-generate detailed and natural actions from the estimated results. In this paper, we describe our approach, Real-time Human Proxy, especially on representing human actions. Also we present experimental results.

1 Introduction

There are several researches on virtual environments for distant interaction. In these researches, a 3-D virtual space is reconstructed, in which each participant is represented as an avatar by computer graphics techniques. Through the reconstructed virtual space, each participant sees and hears other participants' activities from the position where his/her avatar is represented. Therefore, it is called avatar-based interaction.

In avatar-based interaction, an avatar is expected to reflect activities of a participant into a virtual space as if he/she were there. Nevertheless, legacy input devices, such as a keyboard and a mouse, are not sufficient to acquire participant's activities in aspects of quality and quantity. Using such devices, a participant has to intentionally keep feeding their own activities into a system by hand, and acquired information may be neither precise nor rich. To solve this problem, as an input device, we use a vision-based motion capture system (MCS)[1]. Using the MCS, we can acquire rich information of participants without compelling them to do annoying operations.

According to the above idea, we have proposed a concept of Real-time Human Proxy (RHP), which acquires, symbolizes, transfers and represents human information for avatar-based interaction[2]. As the first step of RHP, we focus on nonverbal, or body movement information of humans. In this paper, we discuss Real-time Human Proxy, especially presentation of human action information, i.e., an avatar generation mechanism of RHP.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 883–889, 2005.

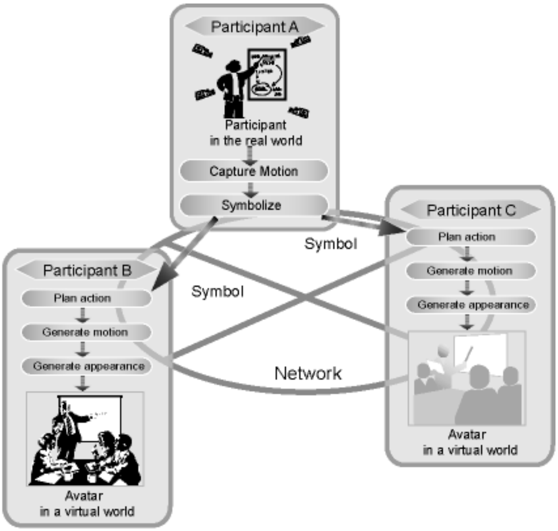


Fig. 1. The concept of RHP

2 Real-Time Human Proxy

Real-time Human Proxy (RHP) is a new concept for avatar-based interaction, which makes avatars act more meaningfully, or expressively, referring to action information acquired by a motion capture system. As the first step, we currently focus on acquisition and representation of human action, or nonverbal information. In the acquisition process, we symbolize the human action information under a given communication environment, such as classroom. In the presentation process, the symbols acquired are presented, or visualized, which are augmented based on the knowledge of the environment. Figure 1 shows a concept of RHP.

The important considerations behind the symbolization are summarized as follows:

- The important aspect of avatar-based communication is that an avatar, or an appearance of a human, can be changed depending on the purpose of communication, attendance, etc. However, only with raw data of human motion, such as motion vectors of body parts, which are acquired by a motion capture system, only an avatar with the same physique as an observed human can be presented. It is quite difficult to present avatars with different physique or avatars with different body structure.
- By a motion capture system, very detailed motion information can not be extracted such as hand postures, face expression at the same time. Such details often express intention and are important for communication. Therefore, here, we interpret, with the aid of knowledge of the purpose communication, limited motion information into intentions of communication, i.e., *symbols*. In presenting the symbols, we can visualize an avatar so as to express the intentions efficiently, i.e., generate detailed motions which are not acquired by a motion capture system.
- The symbolization is also quite helpful to compress the amount of data transfer and to improve QoS (Quality of Service).

Symbolization. On RHP, we acquire human actions instead of human motions. We categorize motion sequences into pre-defined actions, expressing them as symbols. Each symbol is formed by a label of an action and its parameters, such as "walking (p_x, p_y, ν_x, ν_y)" where p_x and p_y are the position, ν_x and ν_y are the velocity of a participant. After recognizing human actions from captured motion data, the system transfers the symbols to the representation side of a virtual space.

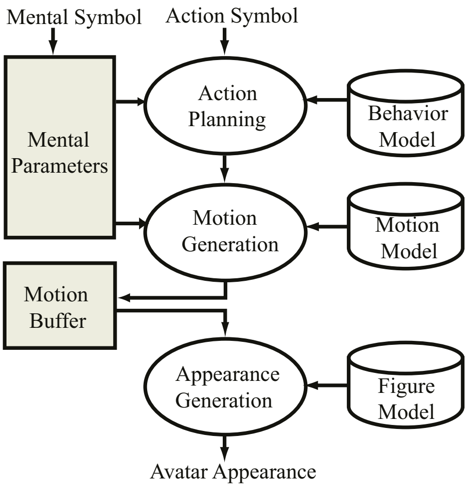


Fig. 2. Process flows

Avatar with Pre-defined Knowledge. We define that an avatar is an object which is participant's substitute in a virtual space. An avatar has pre-defined knowledge to generate its motion and appearance from symbols. But it is time-consuming job to construct or modify the knowledge. Therefore the pre-defined knowledge is to be described in a reusable and extensible form. The details of the knowledge are described in section 3.

Representation of Virtual Space. Generated appearance is represented in a virtual space. A participant is able to see the virtual space in which any participants, including him/herself, are represented as avatars.

3 Avatar Generation

As described in the previous section, RHP allows avatars to be designed beyond constraints of physical structure. To achieve this goal, and to make the avatar generation mechanism general, we have employed a layered structure of the pre-defined knowledge. We divide the pre-defined knowledge into three layers (see Figure 2), and we make them independent as much as possible in order that we can easily modify physical structure of an avatar.

The three layers are *behavior model*, *motion model* and *figure model*. An avatar plans the next action based on *behavior model*, generates a motion corresponding to the next action based on *motion model*, and generates the avatar's appearance with motion based on *figure model*. Here, motion means posture sequences of an avatar's body parts.

3.1 Behavior Model and Action Planning

Action planner generates avatar's next action (action plan) such as "walking" and "raising hand" based on received *symbols*, *behavior model* and *mental parameters*[1]. Action

[1] In this paper, we do not discuss the mental parameters because of page limitation

plans are highly independent of avatar's physical structure. This allows model constructors to modify or replace *behavior model* with taking little care of relations between *behavior model* and other models.

Action Planning. A human can perform multiple actions at the same time if these actions do not require the same body part. For example, "walking (an action using right and left leg)" and "raising hand (an action using right or left arm)" are not mutually exclusive. Therefore, the action planner should generate such multiple actions at the same time. Moreover, on RHP, symbols necessary for interaction depend on the kind of interaction. and it is desirable that *behavior model* can be modified easily, i.e., it should be as simple as possible. From the above considerations, the action planner plans an action referring to *behavior model*, which consists of two kinds of actions; (1)*outward action* is an action transiting from the neutral posture to a specified posture, (2)*Homeward action* is an action transiting from a specified posture to the neutral posture.

The neutral posture is the base posture of starting action. For instance of a human avatar, the posture is a standing posture with his/her arm taking down.

In general, the outward action can be planned in case that the posture of avatar's body parts when a symbol is received is the same as the neutral posture. On the other hand, in case that the posture of avatar's body parts when an action is planned is different from, or collides with, the neutral posture, the action can not be planned. However, if the collided posture is in the homeward action, then it can be planned, it is because avatar's posture is to be the neutral posture soon. A homeward action can be planned after the corresponding outward action was planned.

An action is mainly planned according to a received symbol. However, an avatar often freezes if the avatar acts only when symbols are transmitted, since no symbols are transmitted when a participant does not make any pre-defined actions. Needless to say, such avatar's behavior does not seem natural. To solve this problem, the action planner plans some actions spontaneously such as "folding arms" or "sticking hand into a pocket", which have no influence on interaction. These actions are planned according to the mental parameters. Therefore, during no symbols are transmitted, an avatar can represent actions according to the participant's mental state. To realize it, the system must understand the participant's mental state correctly, which is one of our important future works.

Importance of Action. Each action has a degree of importance for realizing such a function that important actions, or actions according to symbols can be planned more preferentially than others. An example is given below in case when an outward action with a higher degree of importance is selected when an outward action with a lower degree of importance is presented. At first, the homeward action with a lower degree of importance is planned. Then, the outward action with a higher degree of importance is planned immediately. In the opposite case, an action with a lower degree of importance is ignored. Fundamentally, an action according to a symbol is given the highest importance, because the symbol explicitly presents an intention of the participant. On the other hand, an action unrelated to an interaction is given lower importance.

3.2 Motion Model and Motion Generation

There is a motion generator in an avatar which generates the motion based on the planned action, *motion model* and mental parameters. *Motion model* stores detailed motion information corresponding to each planned action.

Motion Generation. *Motion model* is represented as a table of correspondence between an action generated by the action planner and motion information which consists of the following information.

1. Keyframe sequence:$Q_1, Q_2, \cdots, Q_N$
2. The number of frames in the motion:M
3. Frame numbers of keyframes:$p_1, p_2, \cdots, p_N$
4. Interpolation function : $f(i)|i = 0, 1, \cdots, M$

The motion generator generates a motion, or posture sequence, corresponding to a received action. Keyframes expressed with Quaternions are key postures in a motion. Quaternion Q is defined using a rotation axis (V_x, V_y, V_z) and a angle θ as equation (1),

$$Q = (V_x \sin\frac{\theta}{2}, V_y \sin\frac{\theta}{2}, V_z \sin\frac{\theta}{2}, \cos\frac{\theta}{2}). \tag{1}$$

Then, a motion is generated by interpolating between keyframes by using of interpolation function $f(i)$ where i is a frame number in a motion. $f(i)$ is represented in a Bezier function. The process of interpolation between Q_1 and Q_2 is described below. The difference between Q_1 and Q_2, called $Q_{\rm diff}$, is calculated with equations (2), (3), (4) and (5),

$$\begin{aligned} Q &= (x, y, z, w) && (2)\\ \bar{Q} &= (-x, -y, -z, w) && (3)\\ Q_A Q_B &= (v_A \times v_B + w_A v_B + w_B v_A, -v_A \cdot v_B + w_A w_B) && (4)\\ Q_{\rm diff} &= Q_2 \bar{Q}_1, && (5) \end{aligned}$$

where $Q_A = (x_A, y_A, z_A, w_A) = (v_A, w_A)$ and $Q_B = (x_B, y_B, z_B, w_B) = (v_B, w_B)$. $(V_{x\,\rm diff}, V_{y\,\rm diff}, V_{z\,\rm diff})$ and $\theta_{\rm diff}$, which are the rotation axis and the rotation angle between Q_1 and Q_2, can be calculated using equation(1). The motion $Q_{in}(i)$ $(p_1 \le i \le p_2)$ for moving from Q_1 to Q_2 is calculated with the equation(6), using $(V_{x\,\rm diff}, V_{y\,\rm diff}, V_{z\,\rm diff})$, $\theta_{\rm diff}$ and $f(i)$.

$$\begin{aligned} Q_{in}(i) &= (V_{x\,\rm diff} \sin\frac{\theta_{in}}{2}, V_{y\,\rm diff} \sin\frac{\theta_{in}}{2}, V_{z\,\rm diff} \sin\frac{\theta_{in}}{2}, \cos\frac{\theta_{in}}{2}) && (6)\\ \theta_{in} &= \theta_{\rm diff} f(i) && (7) \end{aligned}$$

In this example, we describe about a motion using only one body part. In the case of using multiple body parts, it is just to do these operation for every body part. And the motion generator change the interpolation function by moving control points according to the mental parameters. Therefore a motion can be changed according to the mental states at that time.

Motion Buffer. The motion buffer is a kind of queue. Basically, newly generated motions by the motion generator are added to the tail of the motion buffer, and the oldest motion at the head of the motion buffer is removed to generate avatar's appearance by the appearance generator. However, in case that a generated motion does not collide with an already stored motion, the generated motion is not added to the tail but unified motion is stored where the old stored motion was. This is because that these two motions can be represented simultaneously. For example, when a motion "walking (using right and left leg)" is already stored and a motion "pointing with right finger (using right arm)" is generated, these motions do not collide with each other. So they can be represented simultaneously, then they are unified and turn into one motion "walking pointing with right finger(using right and left leg and right arm)," which is stored where the motion "walking" was.

Motion Fusion. As described before, we restrict actions planned by the action planner to only two kinds of actions which are outward actions and homeward actions. This can reduce animator's job, and make action planning simple. On the other hand, our aim is avatar-based interaction, so an action according to a symbol has to be represented immediately. Therefore such a system does not meet our aim that can not represent an outward action until a colliding homeward action has finished. Moreover, it is unnatural that all actions start from the neutral posture. To solve this problem, the motion generator fuses a homeward action and an outward action which collide with each other, in concrete, interpolates two motions.

The process of fusing two motions is described below. To make the explanation simple, both motions are single body part motions and the numbers of frames of both motions equal to M Motion of the homeward action is $Q_{h(i)}$ and motion of the outward action is $Q_{o(i)}$ $(0 \leq i \leq M-1)$. Difference between these motions, called $Q_{\text{diff}(i)}$, is calculated with equation (5) using $Q_{h(i)}$ and $\bar{Q}_{o(i)}$. Using $Q_{\text{diff}(i)}$, rotation axis $(V_{(i)_{x\,\text{diff}}}, V_{(i)_{y\,\text{diff}}}, V_{(i)_{z\,\text{diff}}})$ and angle $\theta_{\text{diff}(i)}$ for moving from $Q_{o(i)}$ to $Q_{h(i)}$ are calculated with equation(1), and the fused motion, called $Q_{c(i)}$, is calculated with equation(6) using an interpolation function $f(i)$. The interpolation function is important in order to fuse two motions smoothly. We have succeeded in obtaining results like Figure 3 using a linear function $f(i) = i/(M-1)$. In case of using multiple body parts, it is just to do these operations for every body part.

3.3 Figure Model and Appearance Generation

Figure model stores avatar's geometry data and physical structure. The appearance generator generates avatar's appearance using the posture from the head of the motion buffer and figure model.

4 Conclusion

In this paper, we propose a concept of real-time human proxy for avatar-based interaction systems, especially we describe the details of avatar generation. According to the new method, we can easily construct the pre-defined knowledge keeping avatar's behavior natural.

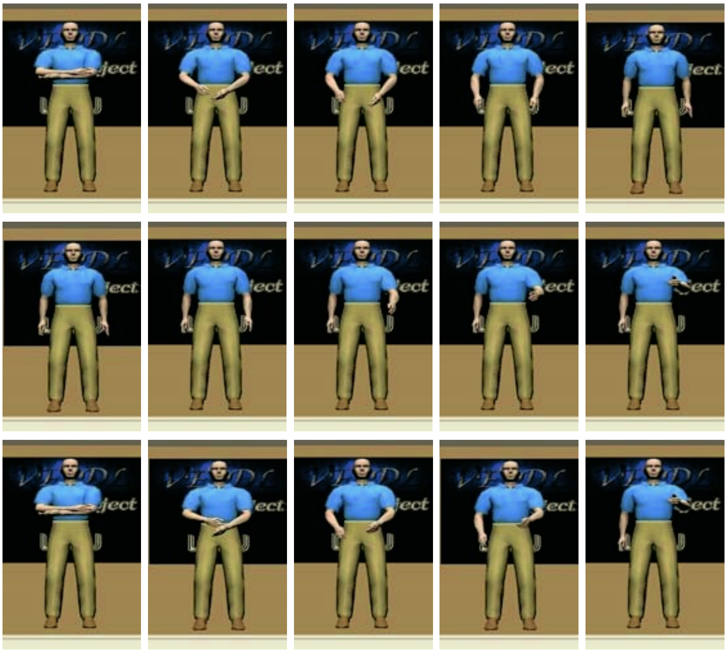


Fig. 3. Fusing Motion: the motion of a homeward action, that of an outward action and the fused motion are shown in the top row, the middle one and the bottom one respectively

Acknowledgment

This work has been partly supported by “Intelligent Media Technology for Supporting Natural Communication between People” project (13GS0003, Grant-in-Aid for Creative Scientific Research, the Japan Society for the Promotion of Science) and “Realtime Human Proxy for Avatar-based Distant Communication” (16700108, Grant-in-Aid for Young Scientists, the Japan Society for the Promotion of Science).

References

1. Date, N., Yoshimoto, H., Arita, D., Yonemoto, S., Taniguchi, R.: Performance evaluation of vision-based real-time motion capture. In: Proc. of Workshop on Parallel and Distributed Computing in Image Processing, Video Processing, and Multimedia, in IPDPS CD-Rom Proceedings. (2003)
2. Arita, D., Yoshimatsu, H., Hayama, D., Kunita, M., Taniguchi, R.: Real-time human proxy: An avatar-based interaction system. In: CD-ROM Proc. of International Conference on Multimedia and Expo. (2004)

Analysis and Synthesis of Help-Desk Responses

Yuval Marom and Ingrid Zukerman

School of Computer Science and Software Engineering
Monash University
Clayton, VICTORIA 3800, Australia
{yuvalm,ingrid}@csse.monash.edu.au

Abstract. We present a corpus-based approach for the automatic analysis and synthesis of email responses to help-desk requests. This approach can be used to automatically deal with repetitive requests of low technical content, thus enabling help-desk operators to focus their effort on more difficult requests. We propose a method for extracting high-precision sentences for inclusion in a response, and a measure for predicting the completeness of a planned response. The idea is that complete, high-precision responses may be sent directly to users, while incomplete responses should be passed to operators. Our results show that a small but significant proportion (14%) of our automatically generated responses have a high degree of precision and completeness, and that our measure can reliably predict the completeness of a response.

1 Introduction

Email inquiries sent to help desks are often repetitive, and generally "revolve around a small set of common questions and issues" (`http://customercare.telephonyonline.com/ar/telecom_next_generation_customer/C`). This means that help-desk operators spend most of their time dealing with problems that have been previously addressed. Further, a significant proportion of help-desk responses contain a very low level of technical content, corresponding, for example, to inquires addressed to the wrong group, or insufficient detail provided by the customer about his/her problem. Organizations and clients would therefore benefit if the efforts of human operators were focused on difficult, atypical problems, and an automated process was employed to deal with the easier problems.

In this paper, we present an initial report of our corpus-based approach to achieving this objective. This approach consists of automatically generating responses to users' "easy" requests on the basis of similar responses seen in a corpus of email dialogues (easy requests have a low level of specific technical detail). Our approach is essentially that used for extractive multi-document summarization, in that similar documents (email responses) are first identified, followed by the automatic extraction of important sentences. However, there is an important difference between our task and traditional multi-document summarization. Normally, the inclusion of an irrelevant information item in a summary does not invalidate the summary. In contrast, in our application, a response email that

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 890–897, 2005.

Request:
Return label was not under the shipping tag and I have been waiting nearly two weeks for a label after reporting it not attached to the box.
Complete response:
I apologize for the delay in responding to your issue. Your request for a return airbill has been received and has been sent for processing. Your replacement airbill will be sent to you via email within 24 hours.

Request:
Hi There, I acquired a HP T3000 1.6G/3.2G Colorado Tape Drive from a friend and would like to know how I go about setting it up for use with WinXP. XP does not seem to detect the drive at all. HELP?
Incomplete response:
Thank you for contacting Hewlett-Packard's Customer Care Technical Center. We are only able to assist customers with in warranty products through our email services. At the present time, we have the following number*s* to contact technical support for your out of warranty product. *Please call PHONENUM. This facility will be available from Monday to Friday between 9.00 AM to 5.00 PM. For additional information, please visit the link given below: WEBSITE.*

Fig. 1. Request with a complete response (top) and request with an incomplete response (bottom)

contains even one incongruous sentence may alienate a user. As a result, the responses generated by our system must have very high relevance (often at the expense of completeness).

To generate such responses, we have developed a procedure that selects high-precision sentences from a cluster of similar responses, and a measure that predicts the completeness of the resultant responses from the features of their source cluster. The idea is that high-precision responses with a high predicted degree of completeness may be sent directly to users, while incomplete responses should be passed to an operator. For example, the top part of Figure 1 shows a request and a complete response automatically generated by our system; the bottom part shows a request and an incomplete response (the additional information in the operator's response and the extra plural in "numbers" in our system's response have been italicized).

Our corpus consists of 30000 email dialogues between users and help-desk operators at Hewlett-Packard. These dialogues deal with a variety of user requests, which include requests for technical assistance, inquiries about products, and queries about how to return faulty products or parts. As a first step, we have automatically clustered the corpus according to the subject line of the first email. This process yielded 38 topic-based datasets that contain between 25 and 8000 email dialogues. Owing to time limitations, the procedures described in this paper were applied to approximately 40% of the data.

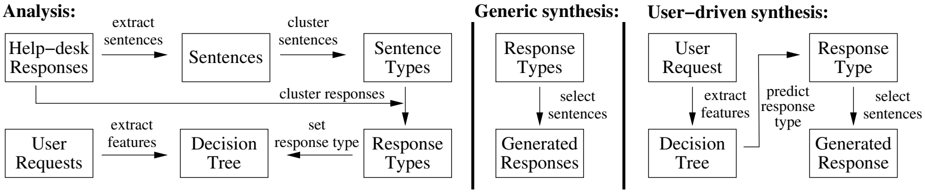


Fig. 2. Analysis and synthesis of responses

2 System Description

Our system analyzes email responses in a particular topic, and then synthesizes responses in two modes: generic or user-driven, as shown in Figure 2. Generic synthesis involves generating model responses, that is, responses that are representative of all the responses seen in the corpus. User-driven synthesis involves generating the most appropriate response for a given user request.

In the **analysis** phase sentences are first extracted from the help-desk responses, and represented by means of binary vectors of size N (number of feature words in the dataset), where element j is 1 if word w_j is present in the sentence, and 0 otherwise. A clustering program is used to group the sentences into *Sentence Types (STs)*. The responses are then represented in terms of the sentence types that they contain. This is achieved by inspecting all the sentences in a response, and assigning a value of 1 for a sentence type if there is a sentence that belongs to this sentence type, and 0 otherwise. This binary representation is used to cluster the responses into *Response Types (RTs)*.

The *Semantic Compactness* (*SemCom*) of each response type is then calculated. This is a measure that predicts whether it is possible to generate a complete, high-precision response from this response type. It calculates, for each response type RT_j, the proportion of highly cohesive and frequent sentence types among the sentence types that have some presence in that response type:

$$SemCom(RT_j) = \frac{\sum_{i=1}^{m}[\ Cohesion(ST_i) > 0.9 \quad \wedge \quad Prop_j(ST_i) > 0.75\ ]}{\sum_{i=1}^{m} Prop_j(ST_i) > 0.1}$$

where $Prop_j(ST_i)$ corresponds to the proportion of responses in RT_j that use sentence type i, m is the number of sentence types discovered in the analysis phase, and $Cohesion(ST_i)$ is a cohesion score calculated for each sentence type:

$$Cohesion(ST_i) = \frac{1}{N}\sum_{k=1}^{N}[\ \Pr(w_k \in ST_i) \leq 0.01 \ \vee \ \Pr(w_k \in ST_i) \geq 0.99\ ]$$

where $\Pr(w_k \in ST_i)$ is the probability that word w_k is used in sentence type i, and N is the number of words in the dataset. The rationale for this cohesion measure is that a cohesive group of sentences should agree strongly on the words they use and the words they omit. Hence, it is easier to find a sentence that adequately represents a cohesive sentence type than a non-cohesive one[1].

Overall, *SemCom* provides a level of confidence that the generated response will be representative of the responses actually used by the help-desk operators. If its value is high, the response type is deemed semantically compact, which means that it is a good candidate for automatic response generation. As the value decreases, so does the confidence of automatically generating a complete response from the response type in question. Before generating a response from a response type, the system compares its semantic compactness with an empirically determined threshold, in order to determine whether an operator should participate in the composition of the reply. In Section 3, we evaluate the semantic compactness measure, and suggest a value for its threshold.

After clustering the responses into response types, a decision tree is trained to predict the response type from features of a user's request. The features currently extracted from the requests are the words that they contain. For each user request in the dataset, the response type is set to the one that the actual response in the corpus belongs to (recall that a response type is a cluster of responses). Thus, user requests are paired with response types and these pairs act as supervised examples for the decision tree.

The **synthesis** phase involves generating a responses from response types. For this purpose we use a modified version of the *adaptive greedy algorithm* proposed in [2] for sentence selection. When selecting sentences for inclusion in a response, the system favours sentences that are representative of sentence types that (a) have a high probability of appearing in the response type in question, and (b) are highly cohesive. The generic synthesis mode involves generating a response for each response type, while the user-driven mode involves generating a single response from the response type predicted by the decision tree for a particular user request.

Examples

The example at the top of Figure 1 is generated from a dataset about product replacements. This dataset contains 1205 email dialogues, and the response emails contain 3598 individual sentences. These sentences are encoded into binary vectors of size 76 (the number of feature words) and clustered into 25 sentence types[2]. Then, the response emails are encoded as vectors of size 25 (the number of sentence types) and clustered, yielding 10 response types. Response type 10, which was used to generate the output in this example, has a perfect semantic compactness (1.0). It represents about 860 responses (71% of the dataset),

[1] The thresholds used in the equations were determined empirically, and chosen specifically to implement a cautious approach that avoids including potentially incongruous sentences in automatically generated responses. However, we have performed a sensitivity analysis which shows that the quality of our responses is largely maintained even if we relax some of these thresholds. For more details on this analysis, see [1]

[2] The clustering program we are using automatically decides on the number of clusters to generate

which all use three highly cohesive sentence types. This means that the sentences shown in the figure are identical to almost all the other sentences in the respective sentence types.

In contrast, the example shown at the bottom of Figure 1 (from a different dataset) is generated from a response type with semantic compactness 0.25. This means that the three highly cohesive and probable sentence types that it uses to generate a response only account for about a quarter (on average) of the sentence types used by the responses represented by this response type. The completing text in the figure shows an example of the kind of sentences that the response type is uncertain about – this kind of information is too specific (phone numbers and operation times).

The user-driven component of the system is currently in development, but we provide here two examples of its preliminary operation. In one dataset the decision tree contains a split on the word "xp", which differentiates two response types. The two response types are very similar, both requesting more information from the user and providing contact numbers for out-of-warranty products. The main difference between them is additional information for XP users, who are referred to another service for additional support. In a different dataset, the responses are so varied that for most of them the system can only generate the sentence "Thank you, HP eServices". However, the decision tree predicts that if the words "cp-2e" or "cp-2w" are present, referring to specific router models, then a response type with very high semantic compactness can be generated, informing the user that he or she has contacted the wrong group and providing the correct address.

3 Evaluation

In this section, we demonstrate the predictive power of our semantic compactness measure, and the ability of our procedure to generate high-precision generic responses with a high level of completeness. Our *SemCom* measure is designed to predict the completeness of an automatically-generated response composed of high-precision sentences. In order to determine the utility of this measure, we examine how well it correlates with the quality of the generated responses.

We assess the quality of a generated response r_g by comparing it with the actual responses in the response type from which r_g was sourced. These comparisons were performed both manually by a panel of human judges, and automatically using three well-known Information Retrieval measures: precision, recall and F-score. Precision gives the proportion of words in r_g that match those in an actual response; recall gives the proportion of words in the actual response that are included in r_g; and F-score is the geometric average of precision and recall. Precision, recall and F-score are then averaged over the responses in r_g's response type to give an overall evaluation of r_g.

Figure 3 shows the relationship between semantic compactness and precision, recall and F-score for the 135 response types created for the different datasets we have used. From the Figure we see that precision is generally high, and is

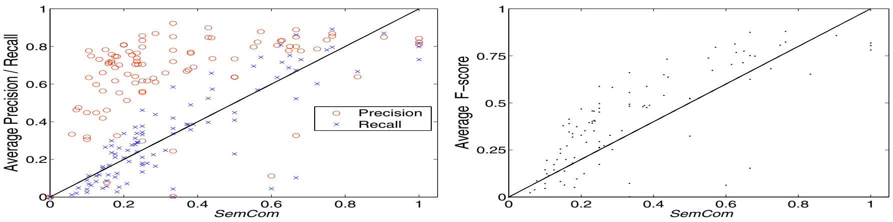


Fig. 3. Relationship between *SemCom* and precision/recall (left), and F-score (right)

uncorrelated with *SemCom*. This is not surprising, as the sentence-selection process is designed to select high-precision sentences. Hence, so long as at least one sentence is selected, the text in the generated response r_g will agree with the text that appears in the responses represented in r_g's response type. In contrast, recall is highly correlated with *SemCom*. A decrease in *SemCom* indicates that fewer sentences are included in the generated response, which therefore covers less of the information in the original responses. As expected from these results, the overall F-score is also highly correlated with semantic compactness (the linear and log correlations between our measure and the F-score are 0.89 and 0.9 respectively). Figure 3 suggests a threshold of 0.7 to indicate high semantic compactness, and a further threshold of 0.4 to indicate medium semantic compactness. The idea is that these thresholds will assist in the selection a response generation strategy for a response type.

As indicated above, we also conducted the following small study in order to assess whether people agree with the predictions made by *SemCom*. We constructed four evaluation sets by selecting four response types with high semantic compactness ($\geq$ 0.7), automatically generating a response from each response type, and selecting 15 actual responses from each response type for comparison[3]. Each evaluation set was given to two judges, who were asked to rate the precision and completeness of the generated response compared to each of the 15 responses in the set. Our judges gave all the automatically generated responses high precision ratings, and completeness ratings which were consistent with our semantic compactness measure.

The overall performance of our system was measured in terms of the proportion of high-precision, complete responses that can be generated from our corpus without human intervention. These are the responses that are represented by response types with high semantic compactness. As mentioned above, Figure 3 suggests a threshold of 0.7 for high semantic compactness. That is, responses that are generated from response types that exceed this threshold could be directly sent to users. This would result in the automatic remittance of approximately 14% of the responses. The application of the medium semantic compactness threshold of 0.4 would result in a further 6% of the generated responses being passed to an operator. The remaining 80% of the responses would

[3] Several of our automatically-generated responses match perfectly the operators' responses. Since these are obvious matches, they were not included in our study

have to be mostly written by an operator. However, this may be a pessimistic estimate, as some response types with a low *SemCom* yield reasonable responses, such as that at the bottom of Figure 1. It is also worth noting that the above percentages vary across the different datasets, which indicates that it may be fruitful to focus the automatic response-generation effort on particular topics.

4 Related Research

The idea of clustering text and then generating a summary from the clusters has been implemented in previous multi-document summarization systems [2–4]. A key issue highlighted in such work is the choice of features used in the clustering. Radev et al. used low-level word-based features [3], while Hatzivassiloglou and colleagues used higher-level, grammatical features obtained through part-of-speech tagging [2, 4].

Our work differs from previous work on clustering and summarization in two respects. Firstly, the high-level features (sentence types) we use to cluster documents are learned from the corpus in an unsupervised manner, using as input only low-level, word-based features. Secondly, our reliance on sentence types enables us to identify response patterns beyond those identified by topic words, and hence allows us to generate different types of summaries within a single topic.

Some examples of user-driven summarization are [5, 6]. The former involves spreading activation from the terms in a query to the terms in news articles to be summarized. The latter involves selecting an answer with the highest posterior probability on the basis of its probability in the corpus and its match with a user's query. The corpus here corresponds to an FAQ, in which there are unique question-answer pairs. The difference between these examples and our user-driven approach is that our system not only matches a user's request with a response, but it can also provide a guarantee of how representative the response is to previous, similar requests in the corpus.

The work that most resembles our approach to automating response generation is [7]. This system retrieves and ranks responses for a query, and then presents a sorted list to a human operator. In this list the most relevant sentences are highlighted, thus assisting the composition of a response. In contrast to our approach, there is no attempt here to fully automate response generation.

5 Conclusion

We have offered a corpus-based approach for the automatic analysis and synthesis of responses to help-desk requests – a task where users exhibit a very low tolerance to irrelevant information. We have also proposed a novel measure that reliably predicts the completeness of a high-precision response, and that can be used to select a response-generation strategy.

Our approach, which uses an unsupervised learning perspective in combination with a simplistic word-based representation, has enabled our system to generate a small but significant proportion (14%) of the email responses in our corpus, without the need for human intervention. We believe that with a more powerful approach, further advances are possible for automating the remaining responses. We are tackling such improvements in our on-going work, which includes performing linguistic analysis to extract higher-level discourse features, and applying machine learning techniques to extract pragmatic features.

References

1. Marom, Y., Zukerman, I.: Corpus-based generation of easy help-desk responses. Technical Report 2005/166, School of Computer Science and Software Engineering, Monash University, Clayton, Australia (2005)
2. Filatova, E., Hatzivassiloglou, V.: Event-based extractive summarization. In: Proceedings of ACL'04 Workshop on Summarization, Barcelona, Spain (2004) 104–111
3. Radev, D., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In: ANLP/NAACL2000 Workshop on Summarization, Seattle, Washington (2000)
4. Hatzivassiloglou, V., Klavans, J., Holcombe, M., Barzilay, R., Kan, M., McKeown, K.: SimFinder: A flexible clustering tool for summarization. In: Proc. NAACL Workshop on Automatic Summarization, Pittsburgh, Pennsylvania (2001)
5. Mani, I., Bloedorn, E.: Multi-document summarization by graph search and matching. In: AAAI97 – Proceedings of the Fourteenth National Conference on Artificial Intelligence, Providence, Rhode Island (1997) 622–628
6. Berger, A., Mittal, V.: Query-relevant summarization using FAQs. In: Proc. 38th Annual Meeting of the Association for Computational Linguistics (ACL2000), Hong Kong (2000) 294–301
7. Carmel, D., Shtalhaim, M., Soffer, A.: eResponder: Electronic question responder. In: CooplS '02: Proceedings of the 7th International Conference on Cooperative Information Systems, Eilat, Israel (2000) 150–161

A Talking Robot and Its Singing Skill Acquisition

Mitsuhiro Nakamura and Hideyuki Sawada

Department of Intelligent Mechanical Systems Engineering, Faculty of Engineering
Kagawa University
2217-20, Hayashi-cho, Takamatsu-city, Kagawa, 761-0396, Japan
sawada@eng.kagawa-u.ac.jp
http://www.eng.kagawa-u.ac.jp/~sawada/index_e.html

Abstract. A talking and singing robot which has equivalent mechanical organs to human vocal system is being developed based on a mechatronics technology under a feedback control. While various ways of vocal sound production have been actively studied so far, a mechanical construction of the vocal system is considered to advantageously realize natural vocalization with its fluid dynamics. Motors are employed for the manipulation of the mechanical system. The robot adaptively learns the relations between motor control parameters and the generated vocal sounds using an auditory feedback learning with neural networks, and sings a song by mimicking a human vocalization. This paper presents the construction of the talking robot and its singing performance, together with the adaptive control for the pitch and phoneme learning. The robot generates vowel and consonant sounds of different pitches by dynamically controlling the vocal cords, vocal tract and nasal cavity.

1 Introduction

Voice is used as primary media in the human communication. It is employed not only in simple daily communication, but also for the logical discussions. Different vocal sounds are generated by the complex movements of vocal organs under the feedback control mechanisms using an auditory system. Vocal sounds and human vocalization mechanisms have been the attractive researching subjects for many researchers so far [1],[2], and computerized voice production and recognition have become the essential technologies in the recent developments of flexible human-machine interface studies.

Various ways and techniques have been reported in the researches of sound production. Algorithmic syntheses have taken the place of analogue circuit syntheses and became widely used techniques [2]-[5]. Sound sampling methods and physical model based syntheses are typical techniques, which are expected to provide different types of realistic vocal sounds [6]. In addition to these algorithmic synthesis techniques, a mechanical approach using a phonetic or vocal model imitating the human vocalization mechanism would be a valuable and notable objective.

Several mechanical constructions of a human vocal system to realize human-like speech have been reported [2],[7],[8]. In most of the researches, however, the mechanical reproductions of the human vocal system were mainly directed by referring to X-ray images and FEM analysis, and the adaptive acquisition of control methods for natural vocalization have not been considered so far. In fact, since the behaviors

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 898–907, 2005.

of vocal organs have not been sufficiently investigated due to the nonlinear factors of fluid dynamics yet to be overcome, the control of mechanical system has often the difficulties to be established.

We are developing a mechanical voice generation system together with its adaptive learning of the control skill for the realization of a talking robot which imitates human vocalization. The fundamental frequency and the spectrum envelope determine the principal characteristics of a sound. The former is the characteristic of a source sound generated by a vibrating object, and the latter is operated by the work of the resonance effects. In vocalization, the vibration of vocal cords generates a source sound, and then the sound wave is led to a vocal tract, which works as a filter to determine the spectrum envelope.

A motor-controlled mechanical model with vocal cords, a vocal tract and a nasal cavity is constructed so far to generate a natural voice imitating a human vocalization [9]-[13]. By introducing an auditory feedback learning with an adaptive control algorithm of pitch and phoneme, the robot is able to autonomously acquire the control method of the mechanical system to produce stable vocal sounds imitating human vocalization skill [10],[11]. This paper describes the adaptive control method of mechanical vocal cords and vocal tract for the realization of a talking and singing robot, together with the singing performance with the use of acquired vocalization skill.

2 Construction of Talking Robot

Human vocal sounds are generated by the relevant operations of vocal organs such as the lung, trachea, vocal cords, vocal tract, nasal cavity, tongue and muscles. The talking robot mainly consists of an air compressor, artificial vocal cords, a resonance tube, a nasal cavity, and a microphone connected to a sound analyzer, which correspond to a lung, vocal cords, a vocal tract, a nasal cavity and an audition of a human, as shown in Figure 1.

An air from the pump is led to the vocal cords via an airflow control valve, which works for the control of the voice volume. The resonance tube is attached to the vocal cords for the modification of resonance characteristics. The nasal cavity is connected to the resonance tube with a sliding valve between them. The sound analyzer plays a role of the auditory system. It realizes the pitch extraction and the analysis of resonance characteristics of generated sounds in real time, which are necessary for the auditory feedback control. The system controller manages the whole system by listening to the generated sounds and calculating motor control commands, based on the auditory feedback control mechanism employing a neural network learning. The relation between the sound characteristics and motor control parameters are stored in the system controller, which are referred to in the generation of speech and singing performance.

2.1 Artificial Vocal Cords and Its Pitch Control

The characteristics of a glottal wave, which determine the pitch and the volume of human voice, are governed by the complex behavior of the vocal cords. It is due to the oscillatory mechanism of human organs consisting of the mucous membrane and muscles excited by the airflow from the lung. Several researching reports about the computer simulations of these movements are available [14], however we have focused on generating the wave using a mechanical model [15].

In this study, we constructed vocal cords with two vibrating cords molded with silicone rubber with the softness of human mucous membrane. Two-layered construction (a hard silicone is inside with the soft coating outside) gave the better resonance characteristics, which is employed in the mechanical voice system. The vibratory actions of the two cords are excited by the airflow led by the tube, and generate a source sound to be resonated in the vocal tract.

The tension of cords can be manipulated by applying tensile force to them. By pulling the cords, the tension increases so that the frequency of the generated sound becomes higher. The relationship between the tensile force and the fundamental frequency of a vocal sound generated by the robot is acquired by the auditory feedback learning before the singing performance, and pitches during the utterance are kept in stable by the adaptive feedback control.

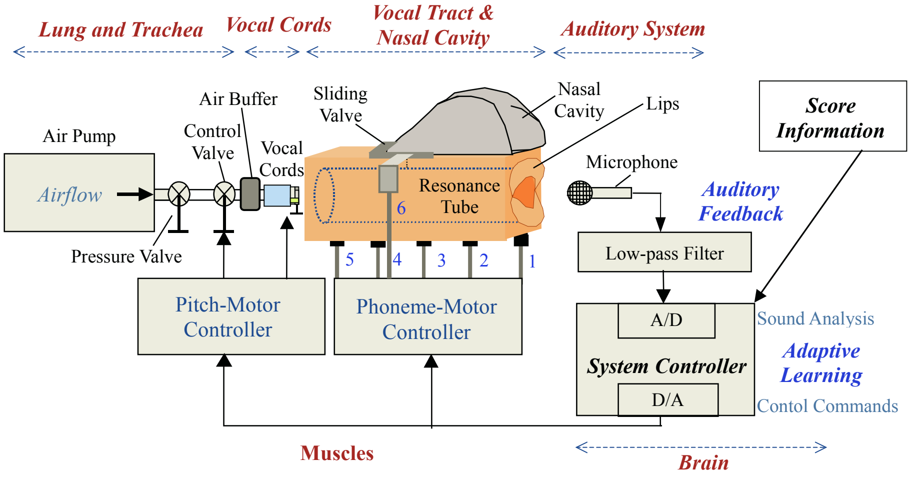


Fig. 1. System configuration

2.2 Construction of Resonance Tube and Nasal Cavity

The human vocal tract is a non-uniform tube about 170mm long in man. Its cross-sectional area varies from 0 to 20cm^2 under the control for vocalization. A nasal tract with a total volume of 60 cm^3 is coupled to the vocal tract. In the mechanical system, a resonance tube as a vocal tract is attached at the sound outlet of the artificial vocal cords. It works as a resonator of a source sound generated by the vocal cords. It is made of a silicone rubber with the length of 180 mm and the diameter of 36mm, which is equal to 10.2cm^2 by the cross-sectional area as shown in Figure 2. The silicone rubber is molded with the softness of human skin, which contributes to the quality of the resonance characteristics. In addition, a nasal cavity made of a plaster is attached to the resonance tube to vocalize nasal sounds like /m/ and /n/.

By actuating displacement forces with stainless bars from the outside, the cross-sectional area of the tube is manipulated so that the resonance characteristics are changed according to the transformations of the inner areas of the resonator. DC motors are placed at 5 positions x_j (j=1-5) from the intake side of the tube to the outlet side, and the displacement forces $P_j(x_j)$ are applied according to the control commands from the motor -phoneme controller.

A sliding valve as a role of the soft palate is settled at the connection of the resonance tube and the nasal cavity for the selection of nasal and normal sounds [13]. For the generation of nasal sounds /n/ and /m/, the sliding valve is open to lead the air into the nasal cavity.

In generating plosive sounds such as /p/, /b/ and /t/, the mechanical system closes the sliding valve not to release the air in the nasal cavity. By closing one point of the vocal tract, air provided from the lung is stopped and compressed in the tract. Then the released air generates plosive consonant sounds like /p/ and /t/.

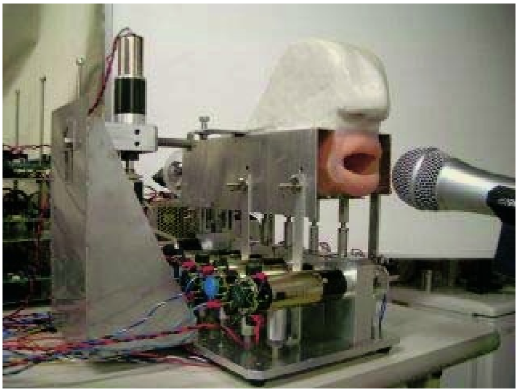


Fig. 2. Structural view of Talking and Singing Robot

3 Learning of Vocalization Skill

An adaptive learning algorithm for the achievement of a talking and singing performance is introduced. The algorithm consists of two phases. First in the learning phase, the system acquires two maps in which the relations between the motor positions and the features of generated voices are described. One is a motor-pitch map, which associates motor positions with fundamental frequencies. It is acquired by comparing the pitches of output sounds with the desired pitches included in melody lines of a music score. The other is a motor-phoneme map, which associates motor positions with phonetic features of the generated voices appeared as lyrics in a score. Then in the performance phase, the robot sings by referring to the obtained maps while pitches and phonemes of produced voices are adaptively maintained by hearing its own outputs.

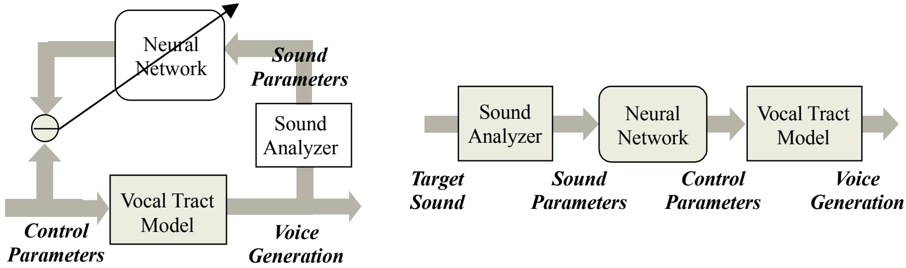
3.1 Adaptive Pitch Learning

The algorithm simulates the pitch learning process of a human in practicing singing [10]. The system starts its action by sending arbitrary values to the pitch controller to let the vocal cord motor and air-control motor move. The pitch of the generated sound is calculated by the sound analyzer of the auditory system, which executes FFT calculations in realtime. The difference between the desired pitch and the current pitch is calculated, and the next motor commands are determined to reduce the pitch difference [15].

As the feedback process is repeated, the pitch difference between the target pitch and the produced pitch decreases. When the pitch difference becomes smaller than a predetermined threshold value, which is currently set to 0.6 Hz, the motor control commands are associated with the target pitch and are stored as the motor-pitch map.

3.2 Learning of Vowel and Consonant Vocalization

The neural network works to associate the sound characteristics with the control parameters of the motors as shown in Figure 3. In the learning process, the network learns the motor control parameters by inputting 10th order MEL cepstrum coefficients derived from sounds vocalized randomly by the robot as teaching signals. The network acquires the relations between sound parameters and the motor-control values of the vocal tract (Figure 3(a)). After the learning, the neural network is connected in series into the vocal tract model as shown in Figure 3 (b). By inputting the sound parameters of desired sounds to the NN, the corresponding form of the vocal tract is obtained.



(a) Learning Phase (b) Running Phase

Fig. 3. Neural Network for Vocalization Learning

In this study, Self-Organizing Neural Network (SONN) was employed for the adaptive learning of vocalization. Figure 4 shows the structure of the SONN consisting of two processes, which are an information memory process and an information recall process. After the SONN learning, the motor-control parameters are adaptively recalled by the stimuli of sounds to be generated.

3.2.1 Information Memory Process

The information memory process is achieved by the self-organizing map (SOM) learning, in which sound parameters are arranged onto a two-dimensional feature map to be related to one another.

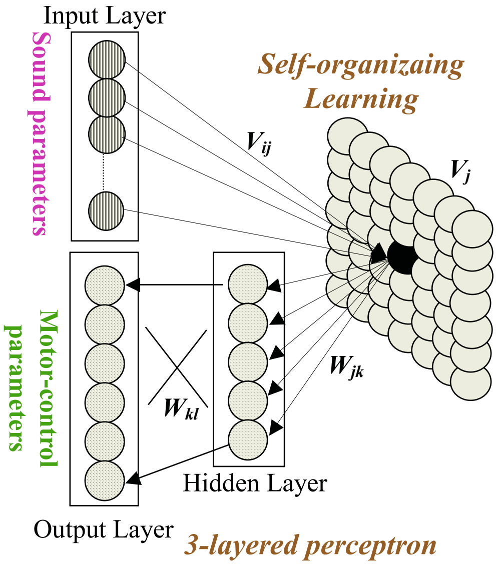


Fig. 4. Structure of Self-Organizing Neural Network (SONN)

Weight vector $\boldsymbol{V}_j$ at node j on the feature map is fully connected to the input nodes x_i [i =1, ... , 10], where 10th order MEL cepstrum coefficients are given. The map learning algorithm updates the weight vectors $\boldsymbol{V}_j$-s. A competitive learning is used, in which the winner c as the output unit with a weight vector closest to the current input vector $\boldsymbol{x}(t)$ is chosen at time t in the learning. By using the winner c, the weight vectors $\boldsymbol{V}_j$-s are updated according to the rule shown below;

$$\begin{gathered}
\boldsymbol{V}_j(t+1)=\boldsymbol{V}_j(t)+h_{cj}(t)\left[\boldsymbol{x}(t)-\boldsymbol{V}_j(t)\right] \\
h_{cj}(t)=\begin{cases} \alpha(t)\cdot\exp\left(-\dfrac{\left\|r_c-r_j\right\|^2}{2\sigma^2(t)}\right) & (i\in N_c\ ) \\ 0 & (i\notin N_c\ ) \end{cases}
\end{gathered} \tag{1}$$

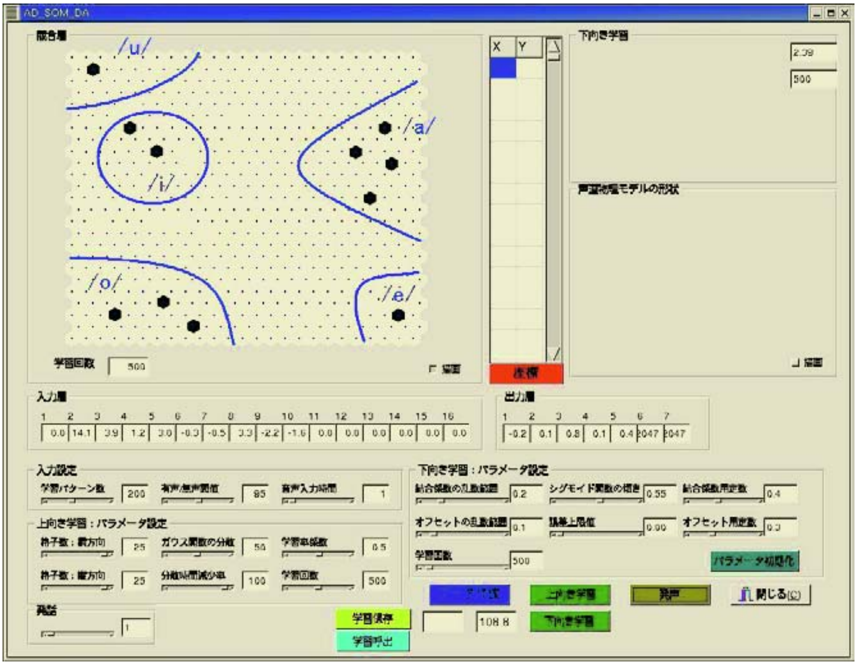
Here, $\| r_c$ - $r_j \|$ is the distance between units c and j in the output array, and N_c is the neighborhood of the node c. $\alpha(t)$ is a learning coefficient which gradually reduces as the learning proceeds. $\sigma(t)$ is also a coefficient which represents the width of the neighborhood area.

In the current system, 20x20 arrayed map $\boldsymbol{V}$ =[V_1 , V_2 , ..., $V_{20\,x20}$] is used, and 200 voices randomly vocalized by the robot are mapped in the learning phase. For testing the mapping ability, Japanese five vowels vocalized by five different people were mapped onto the acquired feature map. As shown in Figure 5, same vowel sounds were mapped close with each other, and five vowels were well categorized according to the differences of phonetic characteristics.

3.2.2 Information Recall Process

Each node in the feature map is associated with motor-control parameters by using the three-layered perceptron. In this study, a conventional back-propagation algorithm was employed for the flexible acquisition of the relation between the shape of the vocal tract and the generated vocal features.

With the integration of the information memory and recall processes, the SONN works to associate sound parameters with motor-control parameters.



学習回数 500

下向き学習: 2.39, 500

入力層

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0.0	14.1	3.9	1.2	3.0	-0.3	-0.5	3.3	-2.2	-1.6	0.0	0.0	0.0	0.0	0.0	0.0

出力層

1	2	3	4	5	6	7
-0.2	0.1	0.8	0.1	0.4	2047	2047

入力設定

学習パターン数 200　有声/無声閾値 85　音声入力時間 1

上向き学習：パラメータ設定

格子数：横方向 25　ガウス関数の分散 50　学習率係数 0.5

格子数：縦方向 25　分散時間減少率 100　学習回数 500

下向き学習：パラメータ設定

結合係数の乱数範囲 0.2　シグモイド関数の傾き 0.55　結合係数用定数 0.4

オフセットの乱数範囲 0.1　誤差上限値 0.00　オフセット用定数 0.3

学習回数 500

発話 1

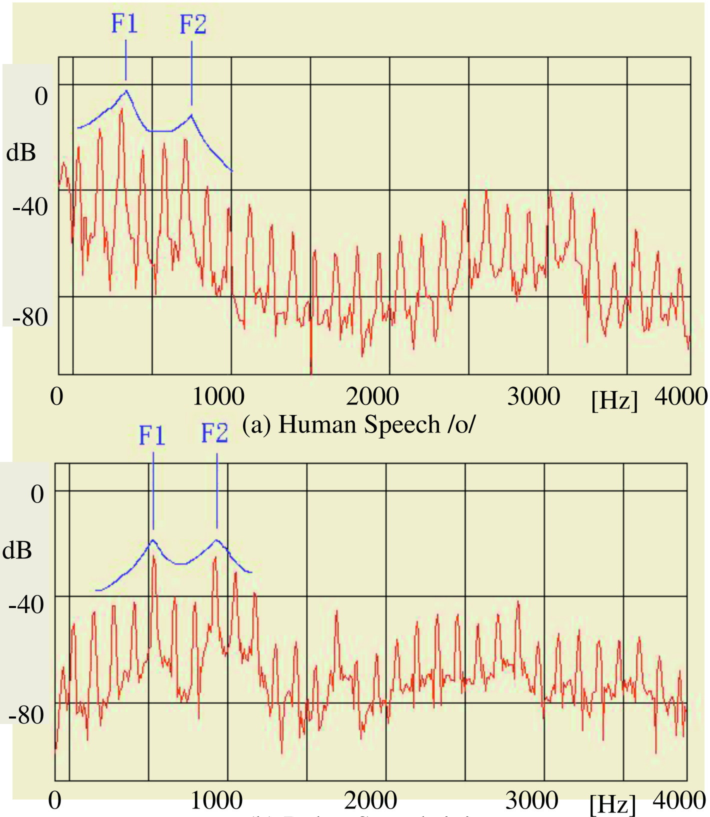
108.8

Fig. 5. An experimental result of Japanese 5 vowel mapping

3.2.3 Experiments for Vocalization Acquisition

After the learning of the relationship between the sound parameters and the motor control parameters, we inputted human voices from microphone to confirm whether the robot could speak by mimicking human vocalization.

Figure 6 shows the comparison of spectra between human /o/ vowel vocalization and robot speech. The first and second formants, which show the characteristics of vowels, were formed as to approximate the human vowels, and the sounds were sufficiently distinguishable. In /a/ vocalization, for example, the glottal side was narrowed while the lip side was open, which was the same way human utter the /a/ sound. In the same manner, features for the /i/ pronunciation were acquired by narrowing the outlet side and opening the glottal side. The experiment also showed the smooth motion of the vocalization. The connection between /a/ and /i/ sounds were well acquired based on the SONN learning, and the /a/ vocalization was transited to /i/ vocalization smoothly by the mechanical system.



(a) Human Speech /o/

(b) Robot Speech /o/

Fig. 6. Comparison of Spectra of Japanese /o/ vowel

4 Singing Performance with Adaptive Control

The singing performance is executed by referring to the acquired two maps, which are the motor-pitch map and the motor-phoneme map. The performance-control manager takes charge of two tasks; one is for the performance execution, and the other is for the adaptive control of pitches and phonemes with the auditory feedback during the performance. Score information is inputted by the use of an interface dialog before the performance. A user selects pitch, duration and lyrics from the lists of musical notations to compose a score, which is stored as score information.

The singing performance is executed according to the performance signals generated by the performance-control manager. The manager has the internal clock and makes a temporal planning of note outputs with the help of the duration information in the score. The score information is translated into motor-control commands by referring to the maps.

During the performance, unexpected changes of air pressure and tensile force cause the fluctuations of sound outputs. The adaptive control with the auditory feedback is introduced by hearing the own output voices. The auditory system observes errors in the pitch and the phoneme so that the system starts fine tuning of produced sounds. The system is able to realize a stable singing performance under the adaptive mechanical control using the auditory feedback.

The system autonomously performs singing with the use of the vocalization skill acquired by the adaptive learning. The robot also makes mimicking performance by listening and following a human singer. The auditory system listens to a human singer and extracts pitch and phonemes from his voice in realtime. Then the pitch and phonemes are translated into motor-control values to let the robot follow the human singer.

5 Conclusions

In this paper a talking and singing robot was introduced, which is constructed mechanically with human-like vocal chords and a vocal tract. By introducing the adaptive learning and controlling of the mechanical model with the auditory feedback, the robot was able to acquire the vocalization skill as a human baby does when he glows up, and generate vocal sounds whose pitches and phonemes are uniquely specified.

A mechanical construction of the human vocal system is considered not only to have advantages to produce natural vocalization rather than algorithmic synthesis methods, but also to provide simulations of human acquisition of speaking and singing skills. Further analyses of the human learning mechanisms will contribute to the realization of a speaking robot, which learns and sings like a human. The proposed approach to the understandings of the human behavior will also open a new research area to the development of a human-machine interface.

Acknowledgment

This work was partly supported by the Grants-in-Aid for Scientific Research, the Japan Society for the Promotion of Science (No. 15700164).

References

1. Y. Hayashi, "Koe To Kotoba No Kagaku", Houmei-do (1979)
2. J. L. Flanagan, "Speech Analysis Synthesis and Perception", Springer-Verlag (1972)
3. X. Rodet and G. Benett, "Synthesis of the Singing Voice", Current Directions in Computer Music Research, PIT Press (1989)
4. K. Hirose, "Current Trends and Future Prospects of Speech Synthesis", Journal of the Acoustical Society of Japan (1992) 39-45
5. Ph. Depalle, G. Garcia and X. Rodet, "A Virtual Castrato", International Computer Music Conference (1994) 357-360
6. J.O. Smith III, "Viewpoints on the History of Digital Synthesis", International Computer Music Conference (1991) 1-10

7. N. Umeda and R. Teranishi, "Phonemic Feature and Vocal Feature -Synthesis of Speech Sounds Using an Acoustic Model of Vocal Tract", Journal of the Acoustical Society of Japan, Vol.22, No.4 (1966) 195-203,
8. K. Nishikawa, H. Takanobu, T. Mochida, M. Honda and A. Takanishi, "Development of a New Human-like Talking Robot Having Advanced Vocal Tract Mechanisms", IEEE/RSJ International Conference on Intelligent Robot and Systems (2003) 1907-1913
9. H. Sawada and S. Hashimoto, "Adaptive Control of a Vocal Chord and Vocal Tract for Computerized Mechanical Singing Instruments", International Computer Music Conference (1996) 444-447
10. H. Sawada and S. Hashimoto, "Mechanical Construction of a Human Vocal System for Singing Voice Production", Advanced Robotics, International Journal of Robotics Society of Japan, Vol.13, No.7 (2000) 647-661
11. H. Sawada and S. Hashimoto, "Mechanical Model of Human Vocal System and Its Control with Auditory Feedback", JSME International Journal, Series C, Vol.43, No.3 (2000) 645-652
12. T. Higashimoto and H. Sawada, "Vocalization Control of a Mechanical Vocal System under the Auditory Feedback", Journal of Robotics and Mechatronics, Vol.14, No.5 (2002) 453-461
13. T. Higashimoto and H. Sawada, "A Mechanical Voice System and its Adaptive Learning for the Mimicry of Human Vocalization", IEEE International Symposium on Computational Intelligence in Robotics and Automation (2003) 1040-1045
14. K. Ishizaka and J. L. Flanagan, "Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Chords", Bell Syst. Tech. J., 50 (1972) 1223-1268
15. T. Higashimoto and H. Sawada: "A Mechanical Voice System: Construction of Vocal Cords and its Pitch Control", International Conference on Intelligent Technologies (2003) 762-768
16. H. Sawada and M. Nakamura: "Mechanical Voice System and its Singing Performance", IEEE/RSJ International Conference on Intelligent Robots and Systems (2004) 1920-1925

Development of a New Vocal Cords Based on Human Biological Structures for Talking Robot

Kotaro Fukui[1], Kazufumi Nishikawa[1,6], Shunsuke Ikeo[1], Eiji Shintaku [1],
Kentaro Takada[1], Hideaki Takanobu[2,4], Masaaki Honda[5], and Atsuo Takanishi[1,2,3]

[1] Department of Mechanical Engineering
Waseda University; 3-4-1 Ookubo, Shinjuku-ku, Tokyo 169-8555, Japan
[2] Humanoid Robotics Institute
Waseda University; 3-4-1 Ookubo, Shinjuku-ku, Tokyo 169-8555, Japan
[3] Advanced Research Institute for Science and Engineering
Waseda University; 3-4-1 Ookubo, Shinjuku-ku, Tokyo 169-8555, Japan
[4] Department of Mechanical Systems Engineering, Kogakuin University, Tokyo, Japan
[5] School of Sport Sciences, Waseda University, Saitama, Japan
[6] Research Fellow of the Japan Society for the Promotion of Science

Abstract. We developed a new talking robot, WT-5 (Waseda Talker No. 5), having novel vocal cords, based on human biological structures. The vocal cords were made from the thermoplastic rubber "Septon", available from Kuraray Co. Ltd. Septon has a similar elasticity to human tissue. The vocal cord model to have a structure similar to the biological structure of the human vocal cords was made. The vocal cords were vibrated like those of a human. This made clean the robot's vowels. With these new mechanisms, the robot could reproduce the human speech in a more biological view and could produce voices nearer to those of a human.

1 Introduction

Because of the importance of speech in human communication, much research has been clarified the mechanisms of human speech. However, because they involve the complexity of aero-acoustics and the movement of speech organs, it is difficult to simulate human speech mechanisms. We have been developing mechanical models of the human speech organs to clarify the mechanisms of speech. We developed talking robots by considering the change in the area of the vocal tract to produce voices like a human.

In 2004, we developed WT-4 (Waseda Talker No. 4) , having human like connectivity between vocal cords and vocal tract, to produce human-like voice, and auditory feedback system to optimize the acoustic parameters of the voices [1]. WT-4 was improved on WT-3 (Waseda Talker No. 3) and could produce all of Japanese vowels and consonant sounds [2]. These parameters, in particular the first and second formants (F1, F2), which are very important parameters in the recognition of vowels, were in the human range.

However, this robot did not prove sufficient for the purpose of clarifying the mechanism of human speech through the modeling of the human vocal organs. Hu-

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 908–914, 2005.

man speech mechanisms comprise very complex movements, and the modeling of these movement mechanisms was untouched in previous robots. And the voices of the previous robots were still different from a human. The difference was caused by the sound source from the vocal cords, and it was not the same as the data estimated data for humans, which attenuated in higher frequency.

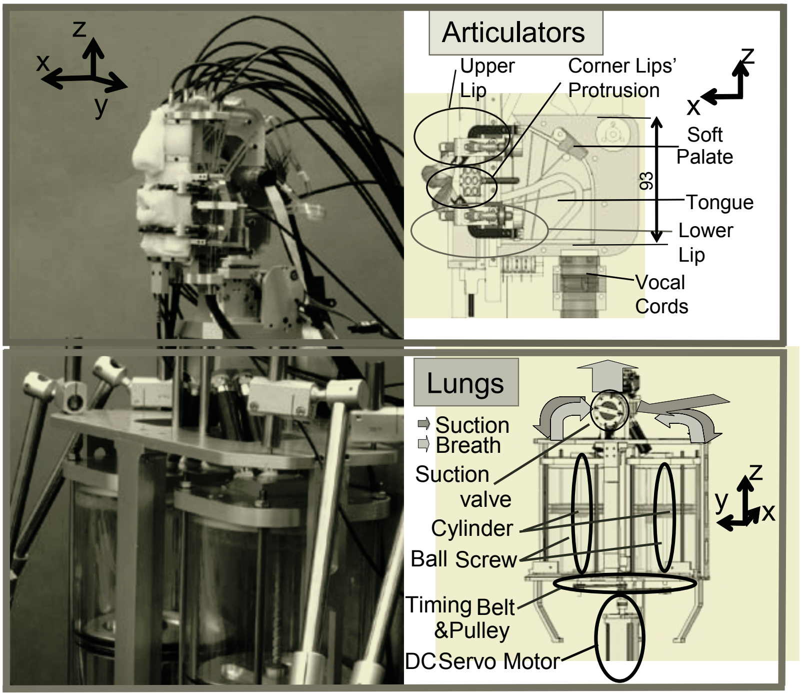


Fig. 1. New anthropomorphic talking robot WT-5

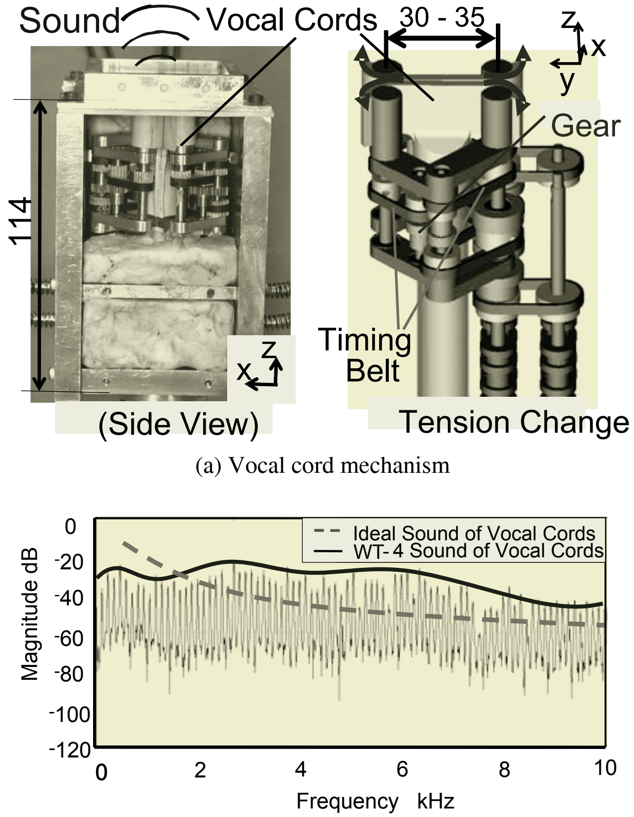
In this paper we describe the development a new vocal cords which was designed to reproduce human speech mechanisms in a more biological manner for a new anthropomorphic talking robot, WT-5, as shown in Fig. 1.

2 Previous Vocal Cords

Human vocal cords are very complex mechanisms comprise muscle and cartilages. We developed mechanical vocal cord models for talking robots, and its vibration was differed from a human's. We needed to develop vocal cords which approximately more to the biological vocal cords of a human to make the sound source of the robot closer to that of a human.

2.1 Mechanism of Human Vocal Cords

The mechanisms of the human vocal cords are mainly changed by cartilages. The thyroid cartilage can change the pitch of the voice through changing the length of glottis. The abduction and adduction of vocal cords can be changed by arytenoid cartilage inside the vocal cords, allowing switching voiced and voiceless sounds [3][4]. Human vocal cords have thin folds and it generates complex vibrations. In the vibration, the lower and the upper edges of the folds were vibrated in different phases.



(a) Vocal cord mechanism

(b) Spectrum of the vocal cords

Fig. 2. WT-4's vocal cords

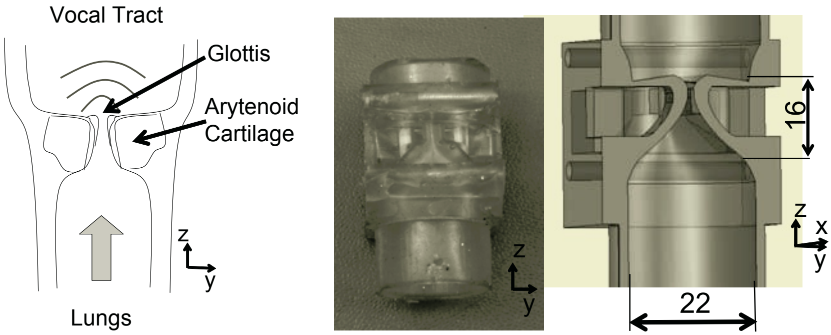
2.2 WT-4's Vocal Cords

WT-4's vocal cords, as shown in Fig. 2(a), consisted of two thin super soft rubber, EPDM [5], having a thickness of 2.0[mm]. The vocal cords had a 4-DOF mechanism having two flexible arms. An independent mechanism rolled around the rubber on each side equally to change pitch. The flexible arms were able to change the glottal length from 30[mm] to 35[mm]. This mechanism could switch between voiced and voiceless sounds, keeping the tension in the vocal cords. Because the vibration of WT-4's vocal cords was very simple, the sound source spectrum was differed from human estimated data, as shown in Fig. 2(b).

3 Forming of Soft Material Based on Human Data

3.1 High Performance Thermoplastic Rubbers

“Septon”, a thermoplastic rubber from Kuraray Co. Ltd [6], was adopted for the new vocal cords. Septon comprises polystyrene and rubber blocks associated in rigid domains. The material can be formed freely and has a greater flexibility than the other elastic materials investigated.



(a) Human (b) WT-5

Fig. 3. Human vocal cords and WT-5’s vocal cords

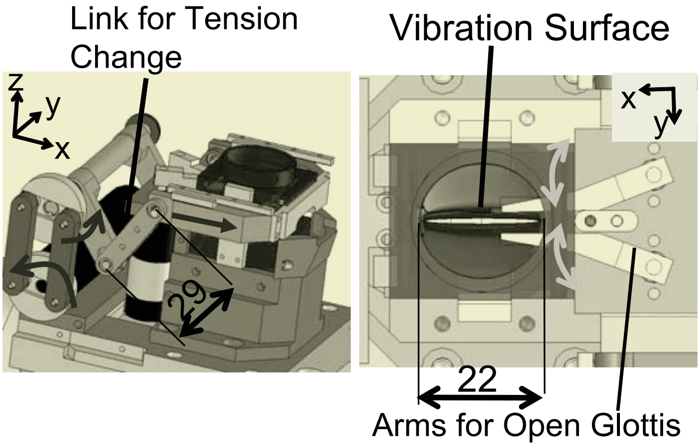


Fig. 4. Vocal cord mechanism of WT-5

3.2 Vocal Cords Mechanism

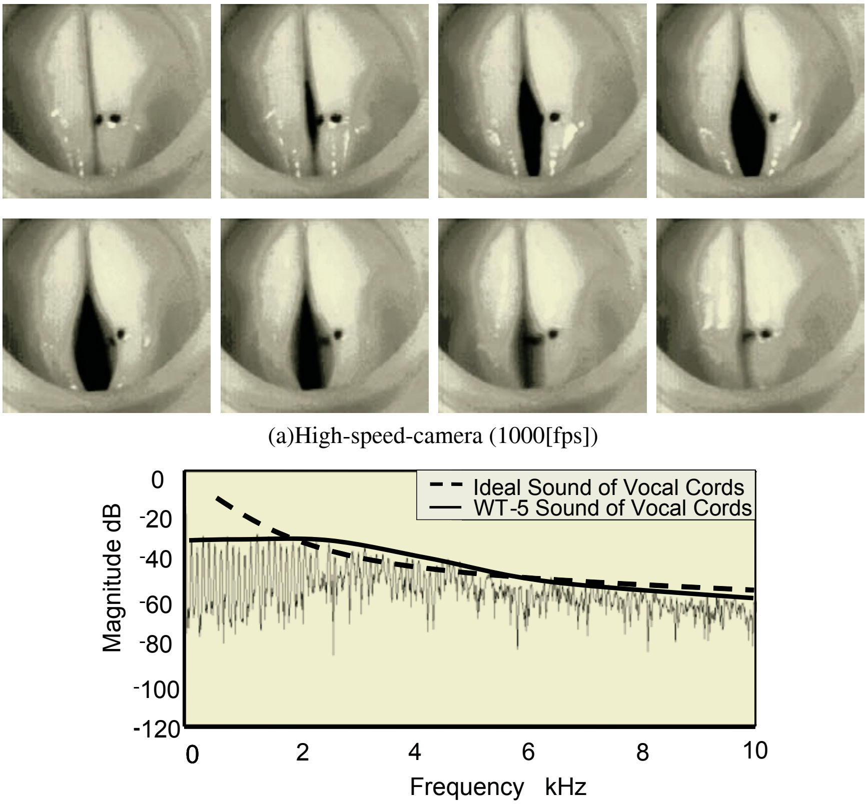
The shape of the new vocal cords was based on that of human vocal cords. The fold shape was mimicked the human’s shape as shown in Fig. 3.

The vocal cords had a 3-DOF mechanism as shown in Fig. 4, 1-DOF to change the tension and 2-DOF to open the glottis. The pitch was changed by the tension of the vocal cords, with the tension being changed by adjusting the length of the vocal cords using the 1-DOF link mechanism. The 2-DOF arm mechanism was used to mimic the abduction and adduction of a human arytenoid cartilage. Switching between voiced and voiceless sounds was achieved by means of arms set on the fold. When the arms adducted, the glottis was closed and vibrated by passing air.

Robot vocal cords were produced based on human biological data. Many types of mold having small variations in the shape parameters of the fold were produced in order to find a suitable sound source.

4 Experiment

We observed the vibration of the new vocal cords using a high-speed-camera. The result was very near to that of a human, as shown in Fig. 5(a). The vocal cords were vibrated by air flow from the lungs of talking robot. The vibration took the form of a wave shape in the vocal cords, as occurs in the vocal cords of a human.


(a)High-speed-camera (1000[fps])

(b)Frequency domain

Fig. 5. Vibration of WT-5's new vocal cords

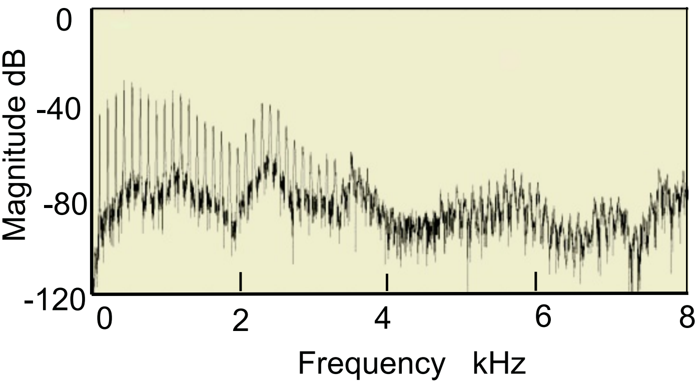


Fig. 6. WT-5's single vowel /a/ attenuated in higher frequencies

In addition, the sound spectrum of new vocal cords was attenuated in higher frequencies that were near to the spectrum of estimated human vocal cords, as shown in Fig. 5(b).

We also experimented the single voice of WT-5 and the voice is attenuated in higher frequencies same as human voice. We show the spectrum of WT-5's /a/ in Fig. 6.

5 Conclusion and Future Work

We developed a new anthropomorphic talking robot, WT-5. The robot had vocal cords based on the biological structures of the human vocal cords. A method of reproducing human vocal cords using elastic materials was developed using the rubber Septon. New vocal cords vibrated in a manner similar to human vocal cords and were made by developing a mold based on data obtained from human vocal cords.

In future work, we intend to develop other vocal organs, besides vocal cords, having more human-like biological forms. Additionally, the controls of the robot become closer to those of a human, to simulate human speech more comprehensively. Through these developments, we aim to clarify the mechanisms of human speech, as well as to develop instruments that will help people with speech difficulties to communicate.

Acknowledgment

This research is supported by Grant-in-aid for Scientific Research (A), 16200015, 2004 by MEXT, Japan. We would like to express our thanks for these people, Dr. K. Honda, Dr. E. Murano and Dr. T. Kitamura on Biophysical Imaging Lab on ATR for the advice of biology of vocal cords, Kuraray Medical Inc. for the supply of Septon and advice on the molding of Septon, SolidWorks Japan K.K. for the provision of 3DCAD software, and Chukoh Chemical Industries for the provision of Teflon coated wire.

References

1. K. Nishikawa, T. Kuwae, H. Takanobu, T. Mochida, M. Honda and A. Takanishi: “Mimicry of Human Speech Sounds using an Anthropomorphic Talking Robot by Auditory Feedback”, Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems. (2004)272-278
2. K. Fukui, K. Nishikawa, T. Kuwae, H. Takanobu, T. Mochida, M. Honda and A. Takanishi: “Development of an Human-like Talking Robot for Human Vocal Mimicry”, Proceedings of the 2005 IEEE International Conference on Robotics and Automation (2005)1449-1454.
3. T. Chiba and M. Kajiyama: “The Vowel -Its Nature and Structure”, Tokyo-Kaiseikan (1942)
4. I. R. Titze: “Principles of Voice Production,” Prentice Hall (1994)
5. EPDM (Ethylene Propylene Diene Monomer) made by TOKYO RUBBER INDUSTRIAL CO., Ltd
6. http://www.septon.info/

An Adaptive Model for Phonetic String Search

Gong Ruibin and Chan Kai Yun

School of Computer Engineering, Nanyang Technological University
Republic of Singapore, Singapore 639798
gongrb@pmail.ntu.edu.sg, askychan@ntu.edu.sg

Abstract. Phonetic string search of written text is an important topic in Information Retrieval (IR). A major difficulty is the inconsistencies between relevance judgements, which makes it possible for a successful method to fail with a new dataset. This paper discusses an adaptive model based on the novel syllable alignment pattern searching algorithm. Experimental results show that it is convenient and effective to be trained for different datasets.

1 Introduction

Phonetic string search of written text is used to identify and retrieve strings that may have similar pronunciation, regardless of their actual spelling. Phonetic string search is useful in Cross-Lingual and Multilingual IR systems, Machine Translation, Transliteration, Back-Transliteration, Spell Check and especially in solving the name string searching problem.

The motivation for this paper comes from our pervious research work [3], in which we proposed a novel algorithm called *Syllable Alignment Pattern Searching (SAPS)*. The experiment showed confusing results with the two different test datasets: while SAPS is obviously the best with the *COMPLETE* dataset [7], it was disappointing with the *Zobel's* [10].

We believe the contradiction comes from the human factor in the relevance judgement. The algorithms are designed after thorough research and are based on the abstraction of the target language and the human perception of the phonetic strings. Regrettably the fuzzy nature of human judgement may easily ruin the designer's effort in setting up a delicate mechanism to regulate the weights of all kinds of factors involved. Thus it is difficult to compare different methods reliably, since they were built and tested on different bases.

Instead of designing a method running well in all kinds of environments, a pragmatic solution is to design an adaptive algorithm for an optimal performance with different datasets. Such an algorithm could be comparably simple but flexible. This is the major interest of this paper.

2 Related Work

For phonetic string search, the different techniques developed so far can be divided roughly into the following four categories:

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 915–921, 2005.

1. Phonetic Transformation Rules. The algorithms in this category use predetermined transformation rules to convert a string into phonetic codes for a later comparison.

The most famous ones are Soundex [4] and Phonix [2].They assume that one phonetic transformation rule works well under all situations. This oversimplification makes their overall performance very poor although they could fetch some good matches that are missed by other algorithms.

In [6], more deliberate consideration have been given to the operations of consonant removal, vowel removal, leading characters and ending sounds in the algorithm called Celko. In [5, 9], necessary spelling-to-sound information is extracted from a dictionary, and is used to produce the most likely phonetic codes for letters or sub strings.

2. Adding Similarity Metric to Soundex. One major weakness in Soundex is that there is no similarity metric for assessing the closeness of two strings - strings are either similar or not similar. Editex [10] adds an edit distance metric with a redefined operation function to the phonetic letter-grouping strategy used by Soundex and Phonix. Editex is easy to deploy and its performance is quite competitive and consistent with different datasets.

3. Assigning Different Weights to Different Positions. It is obvious that the letters in different positions should have different weights in the human judgement of phonetic string similarity, but few researchers have made use of this factor in their algorithms. In Tapering [10], the maximum penalty on the first letter exceeds twice the minimum penalty at the end of the string.

4. Combination of Evidence. In IR, combining the ranked results produced by different retrieval mechanisms could improve the overall performance effectively. The same logic can be applied to phonetic matching, that is, the combination of evidence can lead to a remarkable improvement [6, 7, 10].

3 Syllable Alignment Pattern Searching

The basic idea of SAPS is to segment phonetic strings into syllables and then find the optimal pairing of the corresponding syllables in the two strings. Hence the algorithm is called "Syllable Alignment". The Syllable is "a unit of spoken language consisting of a single uninterrupted sound formed by a vowel, diphthong, or syllabic consonant alone, or by any of these sounds preceded, followed, or surrounded by one or more consonants." [1]. SAPS is based on the following assumptions about how humans recognize phonetic strings,

1. Strings are compared ***syllable-by-syllable*** rather than letter-by-letter.
2. As pitch and stress always fall on the beginning of a syllable, the first letter of the syllable carries the major phonetic information.
3. The Corresponding Syllable Pairs in two strings are located and compared with each other. The Corresponding Syllable Pair have similar spelling/pronunciation, tone and coupled positions in the two strings. E.g., the syllable "son" in the string "Stin**son**" will be compared with the "son" in "Steven**son**", not with "ven" as in the sequence order.

For example, given the following name strings (the 'ˆ' indicates the first letter of a syllable; '-' represents the gap), the preferable alignments is:

$$\begin{array}{ccccccccccccccccccc} \hat{\mathrm{S}} & \hat{\mathrm{t}} & \mathrm{e} & \hat{\mathrm{v}} & \mathrm{e} & \mathrm{n} & \hat{\mathrm{s}} & \mathrm{o} & \mathrm{n} & \quad & \hat{\mathrm{S}} & \hat{\mathrm{t}} & \mathrm{e} & \hat{\mathrm{v}} & \mathrm{e} & \mathrm{n} & \hat{\mathrm{s}} & \mathrm{o} & \mathrm{n} \\ \hat{\mathrm{S}} & \hat{\mathrm{t}} & \mathrm{e} & \hat{\mathrm{v}} & \mathrm{e} & \mathrm{n} & \hat{\mathrm{s}} & \text{-} & \text{-} & \quad & \hat{\mathrm{S}} & \hat{\mathrm{t}} & \mathrm{i} & \text{-} & \text{-} & \mathrm{n} & \hat{\mathrm{s}} & \mathrm{o} & \mathrm{n} \end{array}$$

3.1 Syllable Alignment Pattern Searching

SAPS algorithm consists of three steps: a preprocessing step, a syllable segmentation step and an alignment and similarity-calculation step.

Preprocessing. To deal with some special letter combinations and to simplify the following segmentation step, a preprocessing step is used in SAPS. The design of the preprocessing rule is quite flexible, and can be easily optimized for particular datasets.

Segmentation. This step converts the string into a concatenation of syllables according to the following criteria. E.g., "Stevens"⇒"S|te|ven|s", where '|' represents the boundary between two consecutive syllables.

1. A syllable starts from a consonant, followed by a vowel or a diphthong, and ends before the first letter of the next syllable. E.g. "S|**te**|ven|s".
2. A syllable may end with a consonant if and only if the consonant follows a vowel or a diphthong in the syllable and the consonant precedes another consonant, e.g. "S|te|**ven**|s"
3. A single consonant can be a syllable, e.g. "**S**|te|ven|s".
4. At the beginning (and only at the beginning) of a string, a syllable can start with a vowel, e.g. "**Au**|ty". That is, the beginning of the string is always the start letter of a syllable.

Alignment and Similarity Calculation. A scoring scheme, considering all possible alignments between the first letter of a syllable, normal letter and gap, has been defined as:

- Substitution Function

$$\begin{cases} S(x,y) = \begin{cases} s1, \, x = y \\ s2, \, x \neq y \end{cases} \\ S(\hat{x},y) = \quad s3, \; \hat{x} = y \; or \; \hat{x} \neq y \\ S(\hat{x},\hat{y}) = \begin{cases} s4, \, \hat{x} = \hat{y} \\ s5, \, \hat{x} \neq \hat{y} \end{cases} \end{cases} \tag{1}$$

- Gap Penalty Function

$$\begin{cases} g(x,-) = g1 \\ g(\hat{x},-) = g2 \end{cases} \tag{2}$$

The Substitution Function, $s(x, y)$, indicates the score of aligning a character x with a character y. The Gap Penalty Function, $g(x, -)$ and $g(-, x)$, indicate the cost of aligning a single character x with a gap and aligning a gap with character x. '=' is used to represent two characters that are identical; '≠' to represent two characters that are different (e.g., 'a'='a', 'd'≠'t'). Both the Substitution Function and the Gap Penalty Function are symmetric, i.e., $s(x, y) = s(y, x)$ and $g(x, -) = g(-, x)$. The parameters $s1$–$s5$ and $g1$–$g2$ are the penalty scores of the different operations. Their values depend on how the similarity of the two strings and the different weights for the alignment are defined. In the original SAPS [3], the parameters are set as $s1 = 1; s2 = g1 = -1; s3 = -4; s4 = 6; s5 = -2$; and $g2 = -3$.

For two strings of length m and n, a $(m + 1) \times (n + 1)$ matrix M could be constructed by Dynamic Programming (Eq. 3 and 4) and the value of $M[m, n]$ is the final similarity score of the two strings, where $i \subset [1, m]$ and $j \subset [1, n]$.

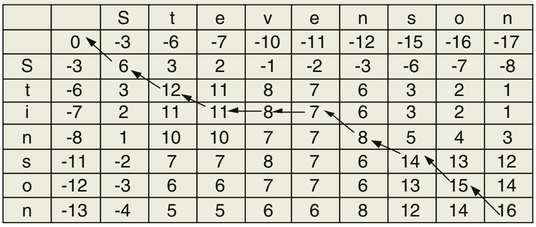
$$M[i, j] = max \begin{cases} M[i-1, j-1] + s(S_1[i], S_2[j]) \\ M[i-1, j] + g(S_1[i], -) \\ M[i, j-1] + g(S_2[j], -) \end{cases} \tag{3}$$

With the initial conditions:

$$\begin{cases} M[0, 0] = 0 \\ M[i, 0] = M[i-1, 0] + g(S_1[i], -) \\ M[0, j] = M[0, j-1] + g(S_2[j], -) \end{cases} \tag{4}$$

To get the actual string alignment, a *TraceBack* procedure is taken from $M[m, n]$ to $M[0, 0]$ in the matrix M to find the path that led to the score. A $(M + 1) \times (N + 1)$ matrix P is used to keep the path information, in which each cell saves a pointer to its parent cell. Then the path from final cell $P[m, n]$ back to $P[0, 0]$ indicates the alignment.

For example, given the two name strings "Stevenson" and "Stinson", their alignment matrix by the original SAPS and traceback path are built as shown in Fig. 1. So the final similarity score is the value of $M[m, n] = 16$, and the corresponding alignment is exactly the one as expected.



		S	t	e	v	e	n	s	o	n
	0	-3	-6	-7	-10	-11	-12	-15	-16	-17
S	-3	6	3	2	-1	-2	-3	-6	-7	-8
t	-6	3	12	11	8	7	6	3	2	1
i	-7	2	11	11	8	7	6	3	2	1
n	-8	1	10	10	7	7	8	5	4	3
s	-11	-2	7	7	8	7	6	14	13	12
o	-12	-3	6	6	7	7	6	13	15	14
n	-13	-4	5	5	6	6	8	12	14	16

Fig. 1. Alignment Matrix of "Stevenson"and "Stinson" and Trace Back Path

4 Adaptive Syllable Alignment Model

The Syllable Alignment Pattern Searching model offers an easy way to adapt to different datasets. Actually there are two possible ways to build an adaptive phonetic string searching algorithm. One is to modify the Phonetic Transforming Rules [6, 10]. The drawbacks of this method are highly demanding of the knowledge of the target language, difficulty in balancing efficiency and complexity, and limited to certain datasets only.

Another method is to train and modify the weights of operations, which is deployed in our algorithm. The original SAPS algorithm could be extended to *Adaptive SAPS* by adapting its parameters $s1$–$s5$ and $g1$–$g2$ in Eq. 1 and 2 to a certain dataset. *Adaptive SAPS* sets $s1 = 1$, $s2 = -1$ and $g1 = -1$ as the basic operation scores, and allow $s4$ to vary between [1,6] and $s3, s5, g2$ between [-6,-1]. The boundaries consider the length of typical English Surname as 4 to 9.

For a given dataset, a small subset and its relevant judgement could be built as the training data. In the training procedure, the *Adaptive SAPS* with all its possible parameter combinations runs on the training data, and the results are compared in the Recall-Precision-Graphs. The parameters of the best result is the trained setting for that dataset.

5 Experiment

5.1 Experimental Environment

Name string search is a typical phonetic string searching problem.Our experiment used the *COMPLETE* surname corpus [7] from Pfeifer and the *Zobel's* dataset [10]. The *COMPLETE* dataset contains 14,972 distinct surnames and 90 randomly selected surnames as queries. Each query has a manually judged set of relevant names. There are 1,187 relevant names in the dataset for the 90 queries. The *Zobel's* dataset has over 30,000 distinct surnames, 100 queries and three sets of judgements: set A on 25 queries, set B on 50 queries and set C on 50 queries.

To deal with the interpolation problem in weak-ordering algorithms, we use the probability of relevance (PRR) method proposed in [8]. For the comparison of the different techniques we use Recall-Precision-Graphs (RPG), in which the x-axis is the *recall* and y-axis is the *probability of relevance.*

5.2 Experiment Results

The 90 queries in COMPLETE dataset were divided into two subsets: 30 queries acted as the training data and the remaining 60 queries as the test data. Firstly the *Adaptive SAPS* ran with the training data and SAPS(-1+2-4-6)— the *Adaptive SAPS* with $s3 = -1$,$s4 = +2$,$s5 = -4$ and $g2 = -6$, was determined as the trained setting. Then SAPS(-1+2-4-6) was compared with Edit Distance, Editex and the original SAPS with the test data. The result showed that *Adaptive SAPS* achieved a further improvement on the original SAPS (Fig. 2).

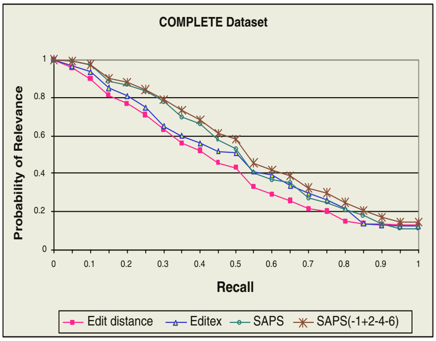
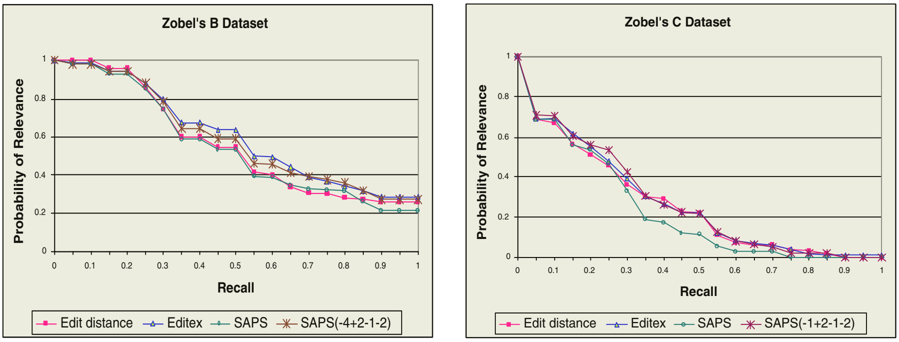


Fig. 2. *Adaptive SAPS* with COMPLETE Dataset



(a) Zobel's B Dataset

(b) Zobel's C Dataset

Fig. 3. *Adaptive SAPS* with Zobel's Dataset

With the Zobel's dataset, we only used the subset B and C, since subset A has only 25 queries, too few for the training and the test. The 50 queries in each subset were divided into 20 training data and 30 test data. In the training session, SAPS(-4+2-1-1) and SAPS(-1+2-1-2) were determined as the final adaptive models for B dataset and C dataset correspondingly and examined in the test session. Fig. 3(a) shows an encouraging result that *Adaptive SAPS* made an obvious improvement on the original SAPS with the B dataset. It was very close with the best performer in this group. The result with the C dataset was even better — the *Adaptive SAPS* became the best in this group as shown in Fig. 3(b).

6 Conclusion

Phonetic string search of written text is an important topic in IR. While the most updated computer intelligence technology is able to supply complex and accurate models, the methods based on simple string similarity comparison are still the best solutions for light-weight search engines, database systems and human-computer interaction interface.

The major difficulty for phonetic string search is the inconsistencies between the relevance judgements, making it impossible to compare different algorithms reliably. More seriously, a successful algorithm developed and tested with one dataset has a high probability of failing with another. The *Adaptive SAPS* proposed in this paper has the advantage to modify its corresponding parameters to the given training data easily and thus achieves optimal performance with the different requirements.Its effectiveness has been tested and results showed its performance is very competitive.

References

1. American-Heritage-Dict. *The American Heritage Dictionary of the English Language, Fourth Edition*. Houghton Mifflin Company, 2000.
2. T. N. Gadd. Phonix: The algorithm. *Program: automated library and information systems*, 24(4):363–366, 1990.
3. Ruibin Gong and Tony Kai Yun Chan. A phonetic string searching algorithm based on syllable alignment. In *Proceedings of the 8th IASTED International Conference on Internet & Multimedia Systems & Applications (IMSA2004)*, pages 108–113, 2004.
4. P.A.V. Hall and G.R. Dowling. Approximate string matching. *Computing Surveys*, 12(4):381–402, 1980.
5. V. Hodge and J. Austin. An evaluation of phonetic spell checkers. Technical Report YCS 338, Department of Computer Science, University of York, 2001.
6. David Holmes and M. Catherine McCabe. Improving precision and recall for soundex retrieval. In *Proceedings of International Conference on Information Technology: Coding and Computing*, Las Vegas, Nevada, April 08-10 2002.
7. Ulrich Pfeiffer, Thomas Poersch, and Norbert Fuhr. Retrieval effectiveness of proper name search methods. *Information Processing & Management*, 32:667–679, 1996.
8. V. V.Raghavan, P.Raghavan, and G. S. Jung. Retrieval system evaluation using recall and precision: Problems and answers. In N. Bellin; C. J. Van Rijsbergen., editor, *Proceedings of the Twelfth Annual International ACMSIGIR Confernce on Research and Development in Information Retrieval*, pages 59–68, New York, 1989. ACM.
9. Liam H. Wugshile and Jos Leubnitz. Using a pronunciation dictionary and phonetic rules for name matching applications. Technical report, Computing Science, RMIT University, 2001.
10. Justin Zobel and Philip Dart. Phonetic string matching: Lessons from information retrieval. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, 1996.

Ontology Modeling and Storage System for Robot Context Understanding

Eric Wang[1], Yong Se Kim[1], Hak Soo Kim[2], Jin Hyun Son[2],
Sanghoon Lee[3], and Il Hong Suh[3]

[1] Creative Design and Intelligent Tutoring Systems Research Center
Sungkyunkwan University, Korea
wang@me.skku.ac.kr, yskim@skku.edu
[2] Department of Computer Science and Engineering, Hanyang University, Korea
{hagsoo,jhs}on@cse.hanyang.ac.kr
[3] Graduate School of Information and Communications, Hanyang University, Korea
shlee@incorl.hanyang.ac.kr, ihsuh@hanyang.ac.kr

Abstract. A mobile robot that interacts with its environment needs a machine-understandable representation of objects and their usages. We present an ontology of objects, with generic shape representations obtained through form-function reasoning. Sets of objects are associated with typical human activities, which supports context understanding. We describe an efficient ontology document storage system, which is based on stable and well-known relational databases. We first design a relational data schema appropriate for Web Ontology Language (OWL) documents, and then develop a transformation mechanism from OWL documents to the relational schema.

1 Object Recognition and Scene Interpretation

To support a robot's interaction with a typical human environment requires a machine-understandable representation of objects, including their shapes, functions, and usages. Object recognition is supported by reasoning from object shape information, while scene understanding is supported by reasoning about sets of objects.

Previous research has explored the relationship between form and function for object recognition. However, these have not directly addressed an ontological representation of objects. The Generic Recognition Using Form and Function (GRUFF) system [1] represents objects as a set of functional elements (mostly planar surfaces), and spatial relations between elements. It performs generic object recognition by matching functional surfaces in the sensor input data to objects' definitions, using customized geometric algorithms. GRUFF organizes object classes into an *is-a* hierarchy, but does not use a standard ontology representation.

Neumann *et al.* [2] performs context-based scene interpretation by modeling scenes as *aggregates*, where an aggregate is a set of entities and their spatial and temporal relations. They represent aggregates of scenes in description logic (DL), and match input models to scene definitions using the RACER DL reasoner [3]. While they do not directly use an ontology representation, their DL representation is essentially equivalent to using the standard OWL ontology language. However, their scene

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 922–929, 2005.

interpretation capability is beyond the current state-of-the-art in description logics, because a complete representation of the relations between entities exceeds the allowed expressiveness of RACER's DL.

2 Object Ontology

We adopt the ontology formalism in developing a generic ontology of objects. We use the standard OWL web ontology language, and the de facto standard Protégé ontology editor with OWL plugin [4]. Using this ontology, we have instantiated a knowledge base of ~300 objects for a typical indoor environment.

Representation of objects. Manufactured objects are typically assembled from multiple components, where each component contributes some specific functionality. Reflecting this, we adopt a hierarchical feature-based representation, shown in Fig. 1. An object is decomposed into a set of features and their spatial relationships, where a *feature* is a functionally significant subset of an object or another feature. Features are characterized by the functions they provide.

Each feature can be further decomposed into more features.

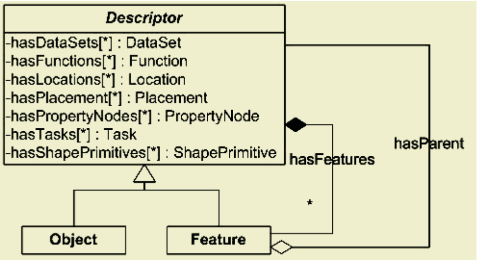

Descriptor
-hasDataSets[*] : DataSet
-hasFunctions[*] : Function
-hasLocations[*] : Location
-hasPlacement[*] : Placement
-hasPropertyNodes[*] : PropertyNode
-hasTasks[*] : Task
-hasShapePrimitives[*] : ShapePrimitive
hasFeatures
hasParent
*
Object
Feature

Fig. 1. Part-whole modeling of Object and Feature classes

Spatial Relations. We define several spatial relations that frequently occur in everyday objects. For each spatial relation, we provide a definition that can be implemented as a (geometric) algorithm. For example, the ***above****(A, B)* relation is defined as: A is above B iff A's highest point is higher than B's highest point (with respect to the gravity direction), and A's lowest point is not lower than B's highest point.

Form-Function Reasoning. We characterize features using generic functions taken from function-based taxonomies for design [5][6]. While a feature is a 3D component, its functional elements, or *organs* [7], may correspond to subsets of its 3D shape. By applying form-function reasoning, we deduce geometric shape requirements for each functional element.

For example, a table's primary function is to limit the downward motion of many objects of any shape. The key feature for a table is a counter, which is typically a thin, rigid 3D slab. A counter's key organ is its top surface. To contact many objects implies many contact points, from which we deduce a planar surface. A table should also minimize the energy required to translate objects to different positions, which implies a horizontal orientation. Hence, we deduce a shape requirement of a *horizontal planar surface* for a counter's top surface.

Geometric Shape Elements. We define a qualitative representation of geometric shape elements, shown in Fig. 2. A shape element has a geometric datum (usually a surface), which represents a generalized portion of a solid's boundary. Other constraints on the allowable orientation, curvature, and tolerance of a shape element are specified using a phrase structure.

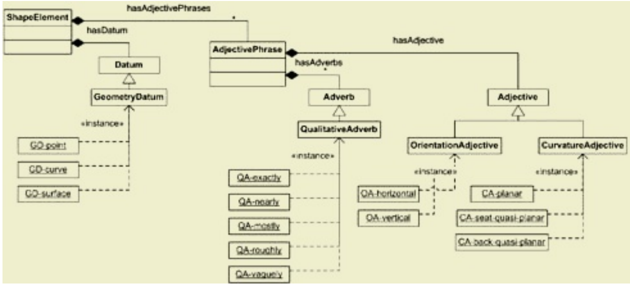


Fig. 2. Representation of geometric shape elements

Representation of Solids. A boundary representation (B-rep) is a 3D model that rigorously describes a solid by enumerating the topological elements of its boundary, including its faces, edges, and vertices. Other solid representations can be converted to B-rep, so a B-rep is a good candidate to be a generic solid representation.

On the other hand, an ontology of objects should also support generic representations of object families. This requires a capability to tolerate wide variations in specific geometry, while capturing the critical geometric relations only.

We adopt a *partial B-rep* scheme, in which a subset of a solid's boundary is fully specified, representing the critical geometric and topological relations only. Remaining portions of the boundary are abstracted away. Each solid has a bounding box data field, reflecting the principle that all real solid objects are bounded. Each feature's shape information is then represented as a partial B-rep with 1 or more geometric shape elements.

3 Use of Ontology for Context Understanding

The object ontology can be used to support object classification. Given a scene with one or more objects:

- Decompose the scene into a set of geometric shape elements, and compute all spatial relations between these shape elements.
- By comparing each shape element in the scene to each feature's required shape elements, and other data such as bounding boxes, classify each shape element into a set of candidate features.
- For each object in the object ontology, check if all of its required features exist, and whether all spatial relations between its features are satisfied. This groups a set of features and spatial relations into a new instance of that object class.
- Repeat using only the unassigned shape elements in the scene data, until all input elements have been assigned to some object.

For scene understanding, we take the perspective that a human activity is characterized by a set of objects that are typically used during that activity. We instantiate associations from sets of objects to activities. Then given a scene of multiple objects, the corresponding activities can be deduced.

4 Ontology Document Management

Nowadays, many researchers have a high interest in the context understanding services, and in reality their concepts have been applied to various application areas, especially such as home networking, telematics, and intelligent robotics. Because the concept of context-awareness is basically considered to be supported by the semantic web technologies, most applications which want to provide context-understanding services may use or adopt semantic web-related international standards: OWL as a web ontology language and Web Services for interaction between software modules [8]. For the efficient support of context understanding services, first of all, the way to store and manage ontology documents should be discussed [9]. In this section, we propose a new data schema based on the relational database and develop a transformation mechanism from OWL ontology documents to the relational schema.

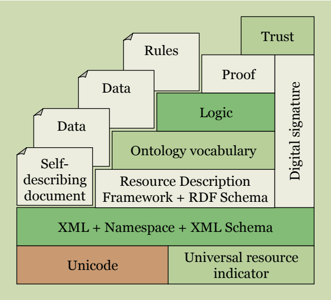

Trust
Rules
Proof
Data
Logic
Data
Ontology vocabulary
Digital signature
Self-describing document
Resource Description Framework + RDF Schema
XML + Namespace + XML Schema
Unicode
Universal resource indicator

Fig. 3. Semantic Web Stack

The XML-based semantic web specification by W3C is composed of five stack layers: resource description framework, ontology vocabulary, logic, proof, and trust layer as in Fig. 3. Here, only resource description framework and ontology vocabulary layers are officially standardized as RDF (RDFS) and OWL, respectively. With RDF and RDFS, we can not only describe web resources with simple statements, but also define classes and properties that may be used to describe other classes and properties as well as web resources. On the other hand, OWL, a revision of the previous DAML+OIL, provides more facilities for expressing meaning and semantics by extending RDF and RDFS. Within these semantic languages, a web resource is represented by a simple statement of a triple data structure (subject, predicate, object). Up to date, several semantic query languages are proposed, for example, such as RQL and RDQL for querying RDF and RDFS documents, DQL for DAML+OIL documents, and OWL-QL for OWL documents. Now, RDQL and OWL-QL are considered as de-facto standards.

As some outstanding toolkits related with the semantic web ontology storage system, Jena [10] and Sesame [11] can be considered to compare with our proposed system. Jena, developed by HP, is a Java-based semantic web framework in which users can easily build semantic web-enabled applications. Using Jena, we can store XML-based semantic web documents including RDF, RDFS, DAML+OIL and OWL, and query the stored documents with RDQL query language. On the other hand, Sesame by NLNet Foundation is basically a storage system for RDF(S) documents, which supports RQL and RDQL to query the stored documents. Being extended by BOR [12], Sesame+BOR can support DAML+OIL documents as well as RDF(S) documents.

Our main contribution in the ontology storage system is to propose an appropriate data schema for ontology documents and develop an ontology document management system in which semantic queries can be processed within the reasonable time even using small amount of data stored.

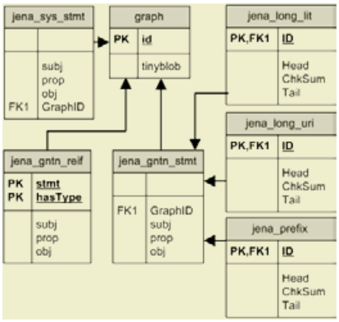


Fig. 4. Jena's Data Schema

5 Proposed Relational Data Schema

Our system's generic goal is to propose a relational schema in which OWL semantic documents can be effectively stored and then efficiently retrieved when querying the stored documents with RDQL. Because DAML+OIL and OWL are almost identical semantic document format, Jena and Sesame+BOR systems which all support DAML+OIL document format and RDQL semantic query language can be compared with our proposed storage system.

In principle, Jena is designed for each document-based storage system. In other words, whenever we insert a semantic document to Jena, two tables, jena_g*n*t*n*_reif and jena_g*n*t*n*_stmt (Here, *n* is any number), dependent on the document are newly generated while other basic common tables are shared as in Fig. 5. In case of the query processing in Jena, we can query only for each document contents while Jena may load into the memory model. On the other hand, Sesame+BOR provides the common relational schema which would be shared for all inserted documents as in Fig. 5. Hence, the query can be targeted to the all data stored in Sesame+BOR system. But, note that Sesame+BOR's table will generally contain too much data because it may generate much more additional information by inferring from the original inserted document. In this regard, we think that our proposed storage system should be like Sesame+BOR but contain as less data as possible without loss of information compared to the original documents. Because of that, we assure that our proposed system be appropriate for embedded systems with the limited resources such as intelligent robotics.

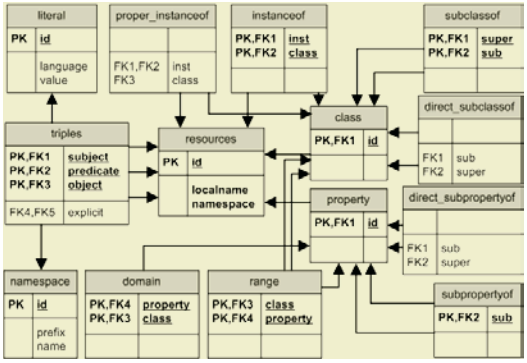


Fig. 5. Sesame+BOR's Data Schema

Because OWL is extended upon RDF and RDFS, its constructs composing of a document are inherited from the RDF(S)'s six core constructs, i.e., rdfs:Class, rdfs:Resource, rdfs:subClassof, rdf:Property, rdf:domain, rdf:range, and rdf:subPropertyof. We design and implement, therefore, our storage system with the relational data schema based on these six core constructs as in Fig.6.

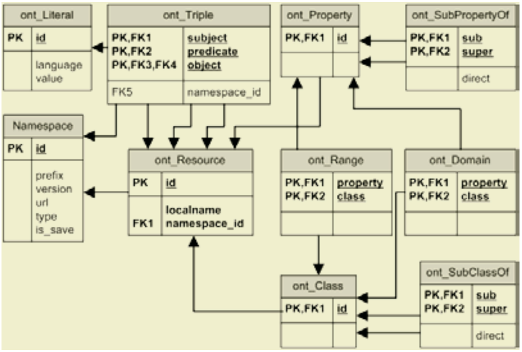


Fig. 6. Our Proposed Data Schema

6 Experiments

To show our system's validness, we compare our system with Sesame+BOR into two different aspects: One is to find the number of records inserted, and the other is to evaluate the average query processing time for the five different semantic queries written by RDQL which are so general in the semantic information processing. As shown in Fig. 7, the number of records generated by Sesame+BOR is as many as twice to our proposed system. If we experiment it with much more big and complex semantic documents, we can find out that our system is more useful for the embedded applications.

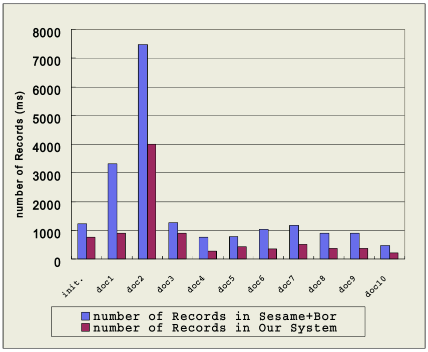


Fig. 7. Comparison in the aspect of the number of records stored

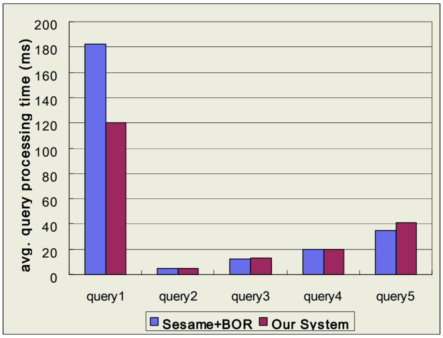


Fig. 8. Comparison in the aspect of the Average Query Processing Time

To compare the two systems in the aspect of the efficiency in RDQL semantic query processing, we have measured the average query processing time for the five different queries as in Fig 8. The five semantic RDQL queries are:

- Query 1: Retrieve all subjects satisfying the predicate and object requirements
- Query 2: Query for the immediate descendant relationship
- Query 3: Query for the descendant relationship
- Query 4: Query on the Range and Domain properties
- Query 5: Retrieve all instances for a certain class

We notice that our system can support the query efficiency similar to Sesame+BOR. As a result, our proposed storage system can provide reasonable query efficiency while keeping small amount of data.

7 Summary and Further Work

We have developed an ontology document storage system based on a relational schema that is appropriate for managing OWL documents in the embedded intelligent robot environment. Some experimental results are additionally presented to justify our proposed storage system.

As further work, we plan to study the way to efficiently process semantic queries, especially using new inference and indexing mechanisms.

Acknowledgement

This work is supported by the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Korea Ministry of Commerce, Industry and Energy.

References

1. L. Stark and K. Bowyer, "Function-based generic recognition for multiple object categories," *Computer Vision, Graphics and Image Processing*, vol. 59, no. 1, pp. 1–21, Jan. 1994.
2. B. Neumann and R. Möller, "On scene interpretation with description logics," *FBI-B-257/04 (Technical Report), Fachbereich Informatik*, Universität Hamburg, 2004.
3. V. Haarslev and R. Möller, "RACER system description," *Proc. Int'l Joint Conf. on Automated Reasoning* (IJCAR 2001), LNAI vol. 2083, Springer, 2001, pp. 701–705.
4. Protégé, http://protege.stanford.edu.
5. C. F. Kirschman and G. M. Fadel, "Classifying functions for mechanical design," *Journal of Mechanical Design*, vol. 120, pp. 475–482, 1998
6. R. B. Stone and K. L. Wood, "Development of a functional basis for design," *Proc. ASME Conf. on Design Theory and Methodology*, Las Vegas, 1999.
7. J. Haudrum, *Creating the Basis for Process Selection in the Design Stage*, Ph.D. Thesis, Institute of Manufacturing Engineering, Technical University of Denmark, 1994.
8. T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*, vol. 284, no. 5, pp. 34–43, 2001.
9. Z. X. Pan and J. Heflin, "DLDB: Extending relational databases to support semantic web queries," *Workshop on Practical and Scalable Web Systems*, pp. 109–113, 2003.
10. J. Carrol and B. McBride, *The Jena Semantic Web Toolkit Public API*, HP Labs, Bristol, 2001, http://jena.sourceforge.net.
11. J. Broekstra, A. Kampman, and F. van Harmelen, "Sesame: A generic architecture for storing and query RDF and RDF schema," Sesame, http://www.openrdf.org.
12. BOR, http://www.ontotext.com/bor/index.html.

Intelligent Two-Way Speech Communication System Between the Technological Device and the Operator

Maciej Majewski and Wojciech Kacalak

Technical University of Koszalin, Department of Mechanical Engineering
Raclawicka 15-17, 75-620 Koszalin, Poland
{maciej.majewski,wojciech.kacalak}@tu.koszalin.pl

Abstract. In this paper there is an intelligent human-machine speech communication system presented, which consists of the intelligent mechanisms of operator identification, word and command recognition, command syntax and result analysis, command safety assessment, technological process supervision as well as operator reaction assessment. In this paper there is also a review of the selected issues on recognition and verification of voice commands in natural language given by the operator of the technological device. A view is offered of the complexity of the recognition process of the operator's words and commands using neural networks made of a few layers of neurons. The paper presents research results of speech recognition and automatic command recognition using artificial neural networks.

1 Intelligent Two-Way Speech Communication

If the operator is identified and authorized by the intelligent speech communication system in Fig. 1, a produced command in continuous speech is recognized by the speech recognition module and processed to the text format [1,3]. Then the recognised text is analysed with the syntax analysis subsystem. The processed command is sent to the word and command recognition modules using artificial neural networks to recognise the command, which next is sent to the effect analysis subsystem for analysing the status corresponding to the hypothetical command execution, consecutively assessing the command correctness, estimating the process state and the technical safety, and also possibly signalling the error caused by the operator. The command is also sent to the safety assessment subsystem for assessing the grade of affiliation of the command to the correct command category and making corrections. The command execution subsystem signalises commands accepted for executing, assessing reactions of the operator, defining new parameters of the process and run directives [2]. The subsystem for voice communication produces voice commands to the operator [4].

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 930–936, 2005.

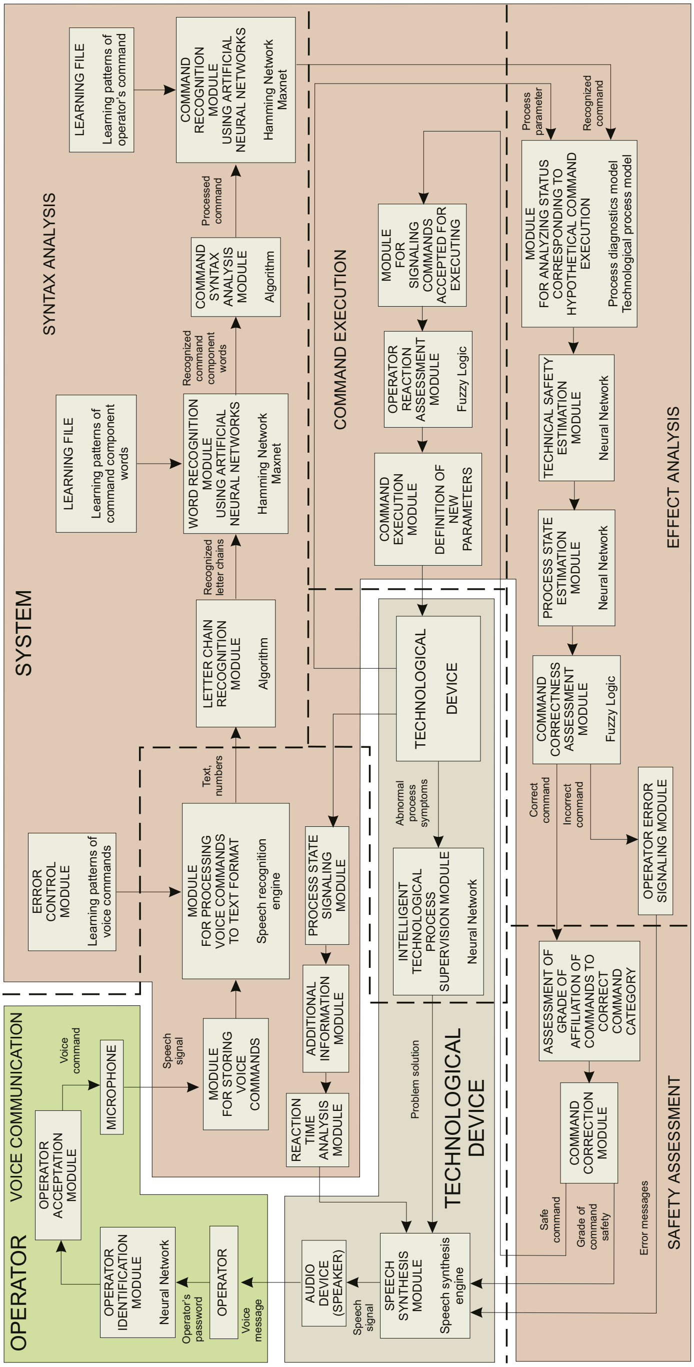

SYSTEM
SYNTAX ANALYSIS
COMMAND EXECUTION
EFFECT ANALYSIS
SAFETY ASSESSMENT
OPERATOR
VOICE COMMUNICATION
TECHNOLOGICAL DEVICE
LEARNING FILE
Learning patterns of operator's command
COMMAND RECOGNITION MODULE USING ARTIFICIAL NEURAL NETWORKS
Hamming Network Maxnet
Processed command
COMMAND SYNTAX ANALYSIS MODULE
Algorithm
Recognized command component words
LEARNING FILE
Learning patterns of command component words
WORD RECOGNITION MODULE USING ARTIFICIAL NEURAL NETWORKS
Hamming Network Maxnet
Recognized letter chains
LETTER CHAIN RECOGNITION MODULE
Algorithm
Text, numbers
ERROR CONTROL MODULE
Learning patterns of voice commands
MODULE FOR PROCESSING VOICE COMMANDS TO TEXT FORMAT
Speech recognition engine
Speech signal
MODULE FOR STORING VOICE COMMANDS
Voice command
MICROPHONE
OPERATOR ACCEPTATION MODULE
OPERATOR IDENTIFICATION MODULE
Neural Network
Operator's password
OPERATOR
Voice message
AUDIO DEVICE (SPEAKER)
Speech signal
SPEECH SYNTHESIS MODULE
Speech synthesis engine
REACTION TIME ANALYSIS MODULE
ADDITIONAL INFORMATION MODULE
PROCESS STATE SIGNALING MODULE
MODULE FOR SIGNALING COMMANDS ACCEPTED FOR EXECUTING
OPERATOR REACTION ASSESSMENT MODULE
Fuzzy Logic
COMMAND EXECUTION MODULE
DEFINITION OF NEW PARAMETERS
TECHNOLOGICAL DEVICE
Abnormal process symptoms
INTELLIGENT TECHNOLOGICAL PROCESS SUPERVISION MODULE
Neural Network
Problem solution
Process parameter
Recognized command
MODULE FOR ANALYZING STATUS CORRESPONDING TO HYPOTHETICAL COMMAND EXECUTION
Process diagnostics model
Technological process model
TECHNICAL SAFETY ESTIMATION MODULE
Neural Network
PROCESS STATE ESTIMATION MODULE
Neural Network
COMMAND CORRECTNESS ASSESSMENT MODULE
Fuzzy Logic
Correct command
Incorrect command
OPERATOR ERROR SIGNALING MODULE
ASSESSMENT OF GRADE OF AFFILIATION OF COMMANDS TO CORRECT COMMAND CATEGORY
COMMAND CORRECTION MODULE
Safe command
Grade of command safety
Error messages

Fig. 1. Architecture of the intelligent two-way speech communication system

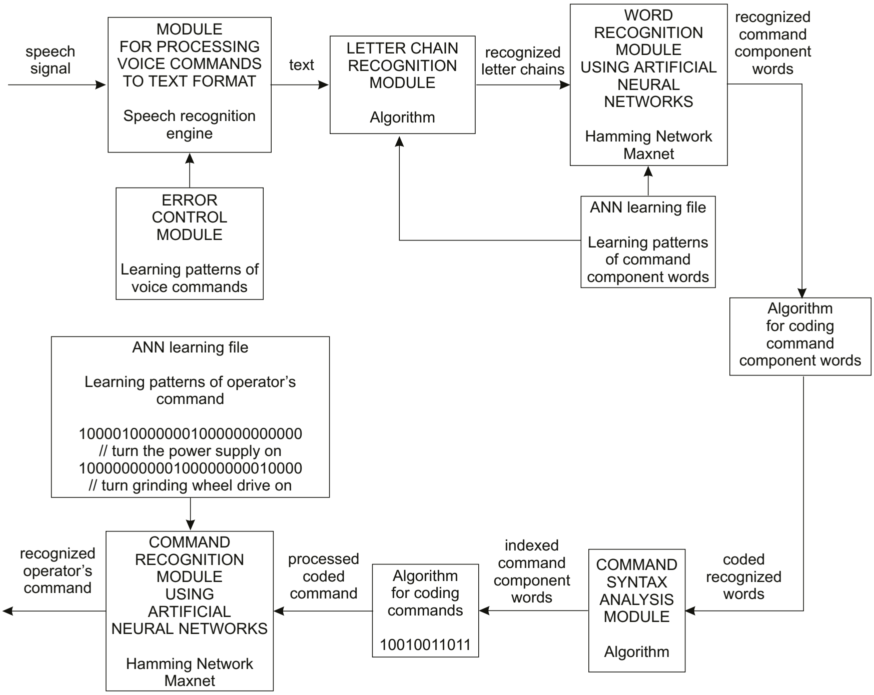

speech signal
MODULE FOR PROCESSING VOICE COMMANDS TO TEXT FORMAT
Speech recognition engine
text
LETTER CHAIN RECOGNITION MODULE
Algorithm
recognized letter chains
WORD RECOGNITION MODULE USING ARTIFICIAL NEURAL NETWORKS
Hamming Network Maxnet
recognized command component words
ERROR CONTROL MODULE
Learning patterns of voice commands
ANN learning file
Learning patterns of command component words
Algorithm for coding command component words
ANN learning file
Learning patterns of operator's command
1000010000000001000000000000
// turn the power supply on
1000000000001000000000010000
// turn grinding wheel drive on
recognized operator's command
COMMAND RECOGNITION MODULE USING ARTIFICIAL NEURAL NETWORKS
Hamming Network Maxnet
processed coded command
Algorithm for coding commands
10010011011
indexed command component words
COMMAND SYNTAX ANALYSIS MODULE
Algorithm
coded recognized words

Fig. 2. Scheme of the automatic command recognition system

2 Automatic Command Recognition and Verification

In the automatic command recognition system as shown in Fig. 2, the speech signal is processed to text and numeric values with the module for processing voice commands to text format. The speech recognition engine is a continuous density mixture Gaussian Hidden Markov Model system which uses vector quantization for speeding up the Euclidean distance calculation for probability estimation. The separated words of the text are the input signals of the neural network for recognizing words. The network has a training file containing word patterns. The network recognizes words as the operator's command components, which are represented by its neurons. The recognized words are sent to the algorithm for coding words. Next the coded words are transferred to the command syntax analysis module. It is equipped with the algorithm for analysing and indexing words. The module indexes words properly and then they are sent to the algorithm for coding commands. The commands are coded as vectors and they are input signals of the command recognition module using neural network. The module uses the 3-layer Hamming neural network in Fig. 3, either to recognize the operator's command or to produce the information that the command is not recognized. The neural network is equipped with a training file containing patterns of possible operator's commands. There was an algorithm created for

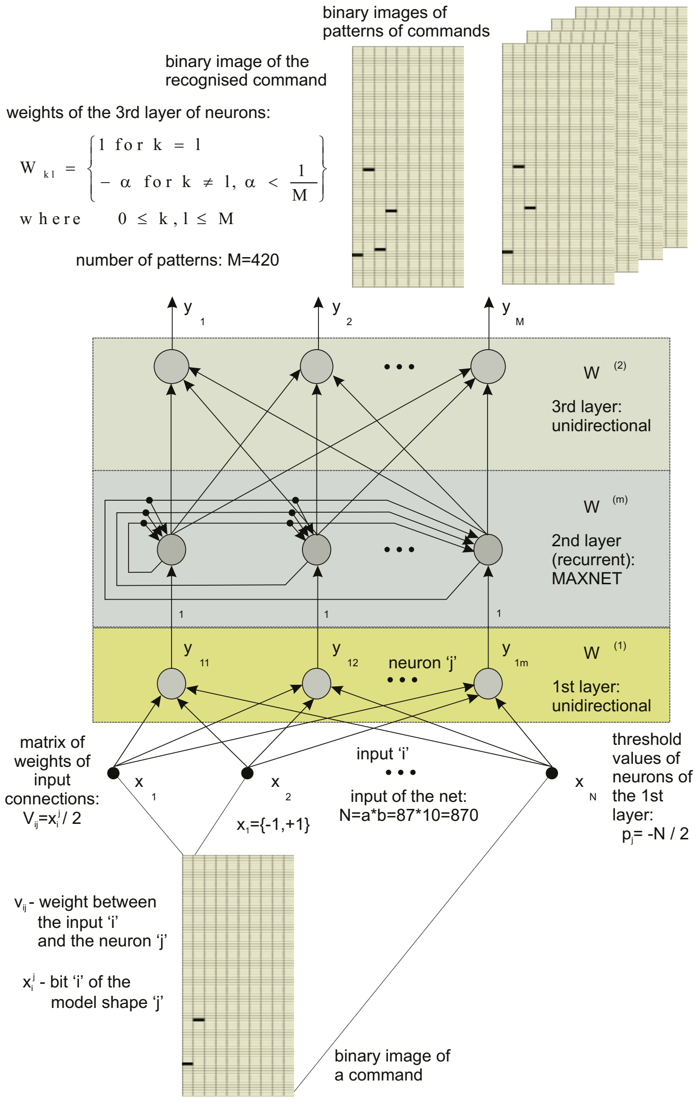

binary images of patterns of commands
binary image of the recognised command
weights of the 3rd layer of neurons:
$W_{kl} = \left\{ \begin{array}{l} 1 \text{ for } k = l \\ -\alpha \text{ for } k \neq l, \alpha < \frac{1}{M} \end{array} \right\}$
where $0 \leq k, l \leq M$
number of patterns: M=420
y_1
y_2
y_M
$W^{(2)}$
3rd layer: unidirectional
$W^{(m)}$
2nd layer (recurrent): MAXNET
1
1
1
y_{11}
y_{12}
neuron 'j'
y_{1m}
$W^{(1)}$
1st layer: unidirectional
matrix of weights of input connections: $V_{ij}=x_i^j / 2$
x_1
x_2
input 'i'
input of the net: N=a*b=87*10=870
x_N
$x_i=\{-1,+1\}$
threshold values of neurons of the 1st layer: $p_j= -N / 2$
v_{ij} - weight between the input 'i' and the neuron 'j'
x_i^j - bit 'i' of the model shape 'j'
binary image of a command

Fig. 3. Scheme of the 3-layer neural network for automatic command recognition

assessing the technological safety of commands. In Fig. 4, the lines present dependence of the force on the grinding process parameters for particular grinding wheels. Basing on the specified criteria, there is the grinding force limit determined for each grinding wheel. Basing on the grinding force limit, there is the table speed limit assigned. According to the operator's command, if the increase of the speed makes a speed of the table smaller than the smallest speed determined from the force limit for all the grinding wheels, then the command is safe to be executed.

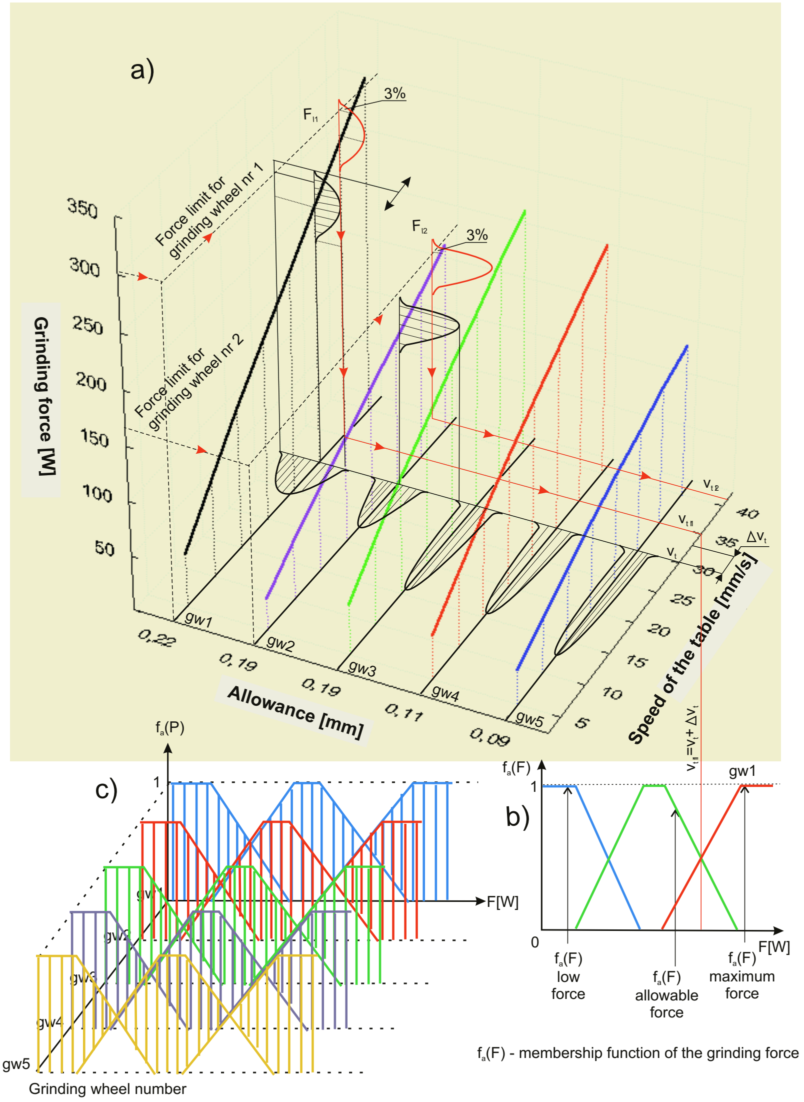


Fig. 4. Algorithm for assessing the technological safety of commands based on the real technological process

3 Research Results of Automatic Command Recognition

As shown in Fig. 5a, the speech recognition module recognizes 85-90% of the operator's words correctly. As more training of the neural networks is done, accuracy rises to around 95%. For the research on command recognition at different noise power, the microphone used by the operator is the headset. As shown in Fig. 5b, the recognition performance is sensitive to background noise. The recog-

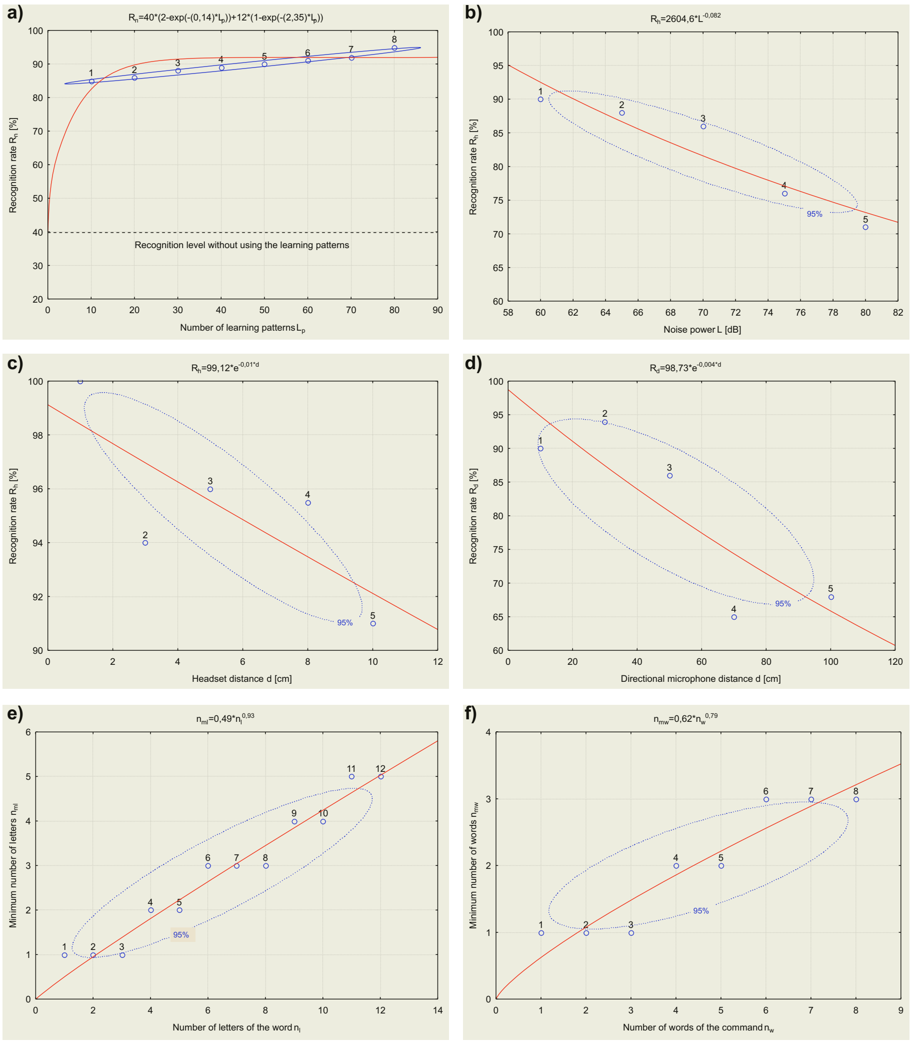


Fig. 5. Speech and command recognition rate

nition rate is about 86% at 70 dB and 71% at 80 dB. Therefore, background noise must be limited while giving the commands. For the research on command recognition at different microphone distances, the microphone used by the operator is the headset. As shown in Fig. 5c, the recognition rate decreases when the headset distance increases. The recognition rate has been dropped for 9% after the headset distance is changed from 1 to 10 cm. Also for the research on command recognition at different microphone distances, the microphone used by the operator is the directional microphone. As shown in Fig. 5d, the recognition rate after 50 cm decreases reaching rate about 65%. As shown in Fig. 5c, the ability of the neural network to recognise the word depends on the number of letters. The neural network requires the minimal number of letters of the word being recognized as its input signals. As shown in Fig. 5d, the ability of the neural network to recognise the command depends on the number of command component words. Depending on the number of component words of the command, the neural network requires the minimal number of words of the given command as its input signals.

4 Conclusions and Perspectives

The condition of the effectiveness of the presented intelligent two-way speech communication system between the technological device and the operator is to equip it with mechanisms of command verification and correctness. In the automated processes of production, the condition for safe communication between the operator and the technological device is analysing the state of the technological device and the process before the command is given and using artificial intelligence for assessment of the technological effects and safety of the command. The research aiming at developing an intelligent layer of two-way voice communication is very difficult, but the prognosis of the technology development and its first use shows a great significance in efficiency of supervision and production humanization.

References

1. Kacalak, W., Majewski, M.: Intelligent Layer of Two-Way Speech Communication of the Technological Device with the Operator, Proceedings of the International Conference on Artificial Intelligence ICAI 2005, 27-30 June 2005, Las Vegas, USA, CSREA, Georgia 2005.
2. Kacalak, W., Majewski, M.: Automatic recognition and safety estimation of voice commands in natural language given by the operator of the technical device using artificial neural networks, Artificial Neural Networks in Engineering ANNIE 2004 Conference, St. Louis, ASME Press, New York 2004, 831-836.
3. Kacalak, W., Majewski, M.: Intelligent Layer of Two-Way Voice Communication of the Technological Device with the Operator, Lectures Notes in Artificial Intelligence 3070, Subseries of Lecture Notes in Computer Science, Springer-Verlag Berlin Heidelberg New York 2004, 610-615.
4. O'Shaughnessy, D.: Speech Communications: Human and Machine, IEEE Press, New York 2000.

Activity-Object Bayesian Networks for Detecting Occluded Objects in Uncertain Indoor Environment

Youn-Suk Song[1], Sung-Bae Cho[1], and Il Hong Suh[2]

[1] Dept. of Computer Science, Yonsei University
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea
corlary@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

[2] The College of Information & Communications, Hanyang University
17 Haengdang-dong, Seongdong-gu, Seoul, 133-791, Korea
ihsuh@hanyang.ac.kr

Abstract. In the field of the service robots, object detection and scene understanding are very important. Conventional methods for object detection are performed with the geometric models, but they have limitations to be used in the uncertain and dynamic environments. This paper proposes a method to predict the probability of target object with Bayesian networks modeled based on activity-object relations. Experiments in indoor office environment show the usefulness of the proposed method for object detection, which produces about 86.5% of accuracy with environments.

1 Introduction

Recently the studies on service robots have been proliferated in many fields [1]. Especially demands of indoor services for elderly people have increased significantly during last decade. For the good service performance, it is very important for the robots to detect objects properly and efficiently. Traditional approaches to object detection is used to utilize only the information in the images so that they are likely to fail in the case that the objects are small or occluded by other objects in indoor environments [2].

In this paper, we propose a hierarchical Bayesian network model of singly-connected structure called activity-object Bayesian network for reasoning the probability of objects from the discovered objects as evidence. The proposed hierarchical Bayesian networks are efficient to model the relationship of objects because they are more informative than naïve Bayes structure and require much less computational power than multiply, fully-connected structure [3, 4]. For this 'common-cause' structures are adopted as building blocks for efficient design with reusability.

2 Backgrounds

2.1 Related Works

The studies for object detection in the images have a profound history. The traditional approaches are based on the two basic assumptions: "all objects are definable by a

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 937–944, 2005.

relatively small number of explicit shape models" and "all objects have characteristic, locally measurable features" [5]. Under these assumptions, conventional methods use the geometry models to detect locations and directions of the object mainly in industries, but these approaches have limitations in the dynamic environment such as home, office, etc.

There are studies to improve these approaches by the knowledge based approach. Marengoni *et al.* tried to add the reasoning system to Ascender I which is the system to analyze aerial images for detecting buildings. They use hierarchical Bayesian networks and utility theory to select proper visual operator in the given context so that they can improve the cost of computation [6]. Torralba *et al.* proposed a method to recognize the place using Hidden Markov Model with the global vectors collected from images and use them as context information to decide the detection priorities [7]. This approach is useful to make detection more efficient but the errors are inherited from the place recognition systems [7]. In this paper, we propose a sophisticated model to predict the presence of target object more precisely from activity-object relations.

2.2 Bayesian Network

Bayesian network is the DAG (directed acyclic graph) model to evaluate the belief of variables using the dependency between them based on the Bayes' rule. The nodes represent random variables while the edges denote the dependencies of them: parent nodes for cause and child nodes for result. The edge between two nodes makes the joint probability distribution, so that the parent has prior probability $P(p)$ and the child has the conditional probability $P(c|p)$. Using the conditional independency, the joint probability distribution $P(x_1, x_2, x_{3,}..., x_n)$ between the nodes can be factored as follows [8].

$$P(x_1, x_2, ..., x_n) = \prod_i P(x_i \mid Parents(X_i))$$

Each node has values which are mutually exclusive and exhaustive in it. In our model, all nodes take binary values.

3 Activity-Object Bayesian Network

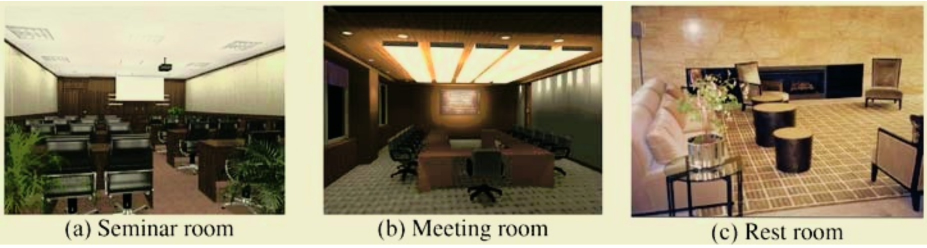
3.1 Service Robot

We construct the activity-object Bayesian networks in the environment having 15 places and 29 objects. The environment is summarized in Table 1 and some of them are shown in Fig. 1.

When the service robot explores in indoor environments for detecting target object, he uses visual processor for recognizing where he is and predicting the presence of objects using place-object Bayesian network for glance-search [9]. After predicting the probabilities, we can decide whether the robot should perform the detailed analysis or not. This approach makes the search more efficient because we can save the costs for detecting the specific objects in images. Activity–object Bayesian networks are used for the detailed analysis for predicting the probability of the target object more precisely from the discovered objects. Activity is more elaborate and efficient classification than place for objects because several activities are occurred in the same place. While the robot does detailed analysis, we can decide again whether it is needed to do more detailed analysis near this place through activity-object Bayesian networks.

Table 1. Service environment

Classification	Contents
Places	Lecture room, Meeting room, Seminar room, Computer room, Prof. office, Admin. office, Guard office, Lab., Hallway, Stair, Hall, Elevator, Toilet, Former Toilet, Rest room
Objects	Table, Side Chair, Castor Chair, Lectern, Cabinet, Bookcase, Garbage can, Sink, Seat toilet, Wall clock, Air conditioner, Telephone, Desktop PC, Notebook PC, Mouse, LCD Monitor, Keyboard, Beam Projector, Projection Screen, Audio, Speaker, Microphone, Wall white board, Castor white board, Partition, Curtain, Water bucket, Door, Window


(a) Seminar room

(b) Meeting room

(c) Rest room

Fig. 1. Some pictures for related place

3.2 Structure of Activity-Object Bayesian Network

The activity-object Bayesian networks are singly-connected structure like tree. They are composed of three kinds of basic nodes: activity node, class node, and primitive node. Activity node is used for a root and class nodes are used for a root of sub-trees. Common-cause structures are used for sub-trees: it is one of the causal relationships between three nodes. It has two nodes that have another node as a parent commonly. This structure allows us to represent the relationship of objects more easily and precisely than simple causal chain because it is possible to assign the parameters to each node for representing the relevancy between them. Related concept on causality is proposed by H. Reichenbach in *The Direction of Time*, 1956. He proposed 'Principle of the Common Cause' as follows: *If two variables are probabilistically dependent, then either one causes the other (directly or indirectly) or they have a common ancestor* [10].

We also use a common-cause structure as a building block, and we can construct whole structure through combining them hierarchically. It is quite useful from viewpoint of design.

The nodes in the activity-object Bayesian networks have binary values. The primitive nodes are able to accept the evidences as inputs only. About this, we will discuss in the next section. In general, the activity node (i.e. root node) has two basic class nodes by design principles: public class node and private class node. Public class node is composed of sub-trees which can be reused in another activity-object Bayesian network. The relationship between the public class node and the activity node is adjusted through the parameters of public class node. Like this way, we can reuse the sub-trees belonged to the public class node for another Bayesian network just after readjusting the parameters of the root node in it. The descendants of private class node are composed of the nodes to have strong relationship with activity node. So the experts must design them with special domain knowledge. Some design principles are summarized in the following.

- Activity node has only two child nodes: public class node and private class node.
- Public class node is composed of the building blocks commonly used in another activity-object Bayesian networks and private class node is composed of the objects which are highly related with activity node.
- The parameter of class nodes means hierarchical relationship and that of primitive nodes means the relevancy between the siblings.

Basic structure of activity-object Bayesian network structure is shown in the Fig. 2.

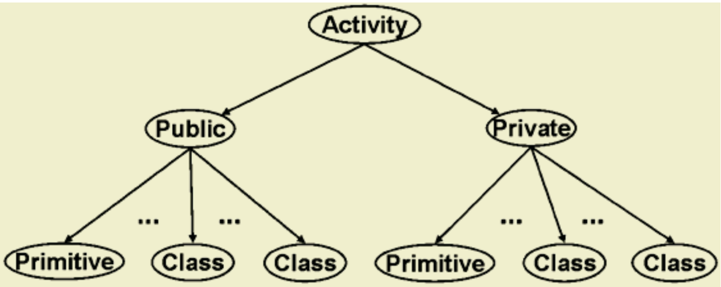


Fig. 2. Basic structure of Activity-Object Bayesian network

- Activity node: Root node. Criterion for making relationship among objects. It has a public and a private class node. Only used for output
- Class node: Root node of sub-tree. It correlates the objects. Its parameter means hierarchical correlation. Only used for output
- Primitive node: Leaf node. It represents objects. Its parameter means horizontal correlation between other primitive nodes in the same building blocks. Input for evidence. Output for probability of objects being present

This hierarchical structure is more informative than naïve Bayesian structures and requires less computational complexity than multiply, fully-connected Bayesian networks. The computational specifications of the structures are compared in Table 2.

Table 2. The complexity of several structures

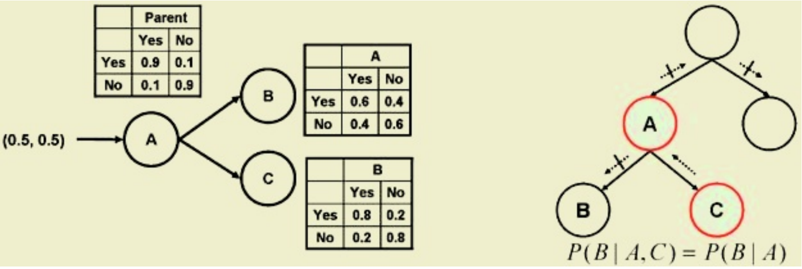
	Naïve	Fully-Conn.	Hierarchical. S.
# of objects	n	n	n
# of value	$2^2(n\text{-}1)+2$	$2(2^n\text{-}1)$	$2^2(n+\alpha\text{-}1)+2$
Complexity	$O(n)$	$O(2^n)$	$O(n)$

3.3 Conditional Probability Tables and D-Separation

It is important to maintain the probability distributions of all nodes to be *P*(*C*=*yes*|*P*)=0.5 and *P*(*C*=*no*|*P*)=0.5 after belief-updating without evidences in our problem. Suppose α represent the probability of *P*(*C*=*yes*|*P*=*yes*) and assign 1-α to *P*(*C*=*yes*|*P*=*no*) then by the following formula with uniformed prior probability of activity node we can maintain the probabilities of all nodes as (0.5, 0.5) in the case that there are no evidences in the network.

$$\begin{aligned} P(C_{yes} \mid P) &= P(C_{yes} \mid P_{yes})P(P_{yes}) + P(C_{yes} \mid P_{no})P(P_{no}) \\ &= \alpha \times 0.5 + (1-\alpha) \times 0.5 \\ &= 0.5 \end{aligned}$$

The examples of parameter settings are shown in Fig. 3(Left).



	Parent	
	Yes	No
Yes	0.9	0.1
No	0.1	0.9

	A	
	Yes	No
Yes	0.6	0.4
No	0.4	0.6

	B	
	Yes	No
Yes	0.8	0.2
No	0.2	0.8

Fig. 3. Left: Example of conditional probability tables. Right: d-separation

Probability propagation is performed via a "flow of information" through the network[8] and there are two types of reasoning direction: diagnostic and predictive reasoning. Diagnostic reasoning infers from children to parents and predictive reasoning infers parents to children. D-separation is the notion that one node blocks the propagation between sets of nodes. In the case of common-cause structure, common-node given the evidence blocks the propagation not only between children but also between his ancestor and descendant (Fig. 3(Right)). To solve this problem, we suppose that the class node can be only used as the common-node. It allows us to be free from the problem of d-separation because class node cannot have evidence as input under the assumption.

4 Experiments and Results

4.1 Experimental Environments

The experiments are performed with a presentation activity-object Bayesian network for 6 places (Computer room, Laboratory, Rest room, Conference room, Seminar room and Guard office). The overall structure of the Bayesian network is summarized in Table 3 and shown in Fig. 4.

The purpose of our experiments is to observe the performance of activity-object Bayesian networks designed by experts for prediction of target object. We assume that the service robot moves from place to place and detects some objects randomly which are in the place. We record the probability values and hitting rate for predicting the probability of beam-projector in each place.

Table 3. The description of presentation activity-object Bayesian network

Node	Names	#
Activity	Presentation	1
Class	Public{Furniture{Basic1},ComputerRelated{Accessory},AudioTool}, Private{PresentationTools}	8
Primitive	Chair, Table, Mouse, Keyboard, LCDMonitor, Computer, Speaker, Audio, Lectern,Microphone x2,Board,ProjectionScreen,BeamProjector	14

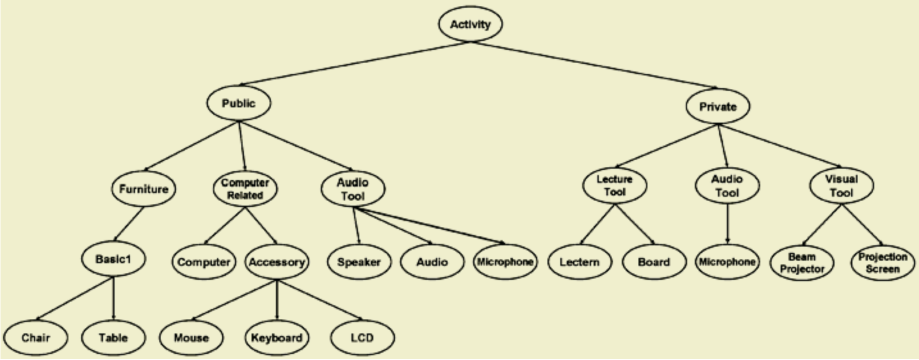


Fig. 4. Overall structure of presentation activity-object Bayesian network

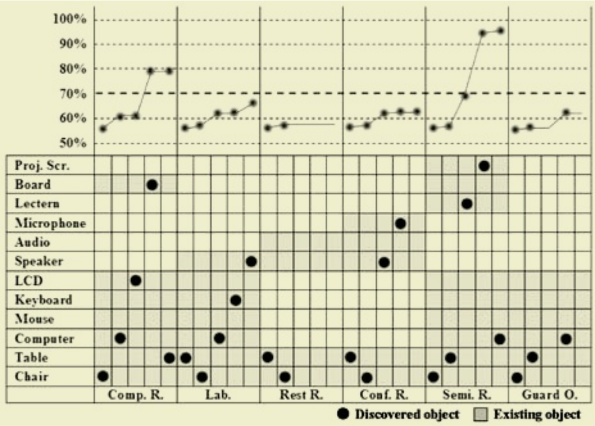


Fig. 5. Probability of beam projector in each place

4.2 Experiments Results

The probability for beam-projector in each place is shown under threshold 70% until the robot finds 5 objects (Fig. 5). Predictions seem reasonable except one case. We can see the robot predicts that a beam projector exists in the computer room and seminar room but not computer room actually. This fact denotes that false-positive error is likely to occur in the similar environment from the result of accumulating evidences. It is important to decide the threshold value and number of times for finding objects for the performance. The hitting rate is summarized in Fig. 6. For this experiment we try to test ten times in each place and try again ten times whole processes. This result also confirms the same fact with the previous one. The overall result is 86.5 %, which shows the proposed method is reliable to predict objects being present.

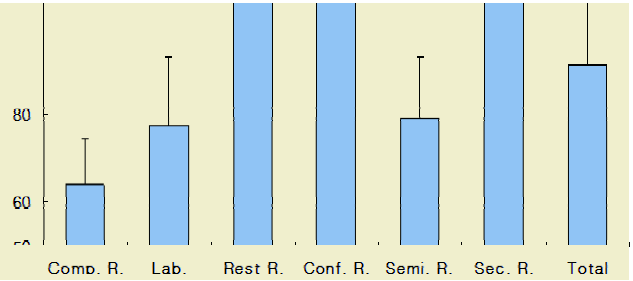


Fig. 6. Hitting rate of predicting the beam-projector

5 Summary and Conclusion

We propose a Bayesian network model for the efficient detection of objects in uncertain environments. Using the hierarchical Bayesian network structure, we can infer the probability of the target object being present. Also, we have tested why this structure is good for our problem including some related issues (posterior probability of nodes without evidences and d-separation) and the useful aspects in terms of design. Our experiments show the reasoning of object being present is helpful for object detection.

In the future work, we will use negative nodes for decreasing the probabilities and try various activity-object Bayesian networks for real robots.

Acknowledgement

This paper was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea.

References

1. P. Dario, *et al.*, "Robot assistants: Applications and evolution," *Robotics and Autonomous Systems*, vol. 18, pp. 225-234, 1996.
2. K. Murphy, *et al.*, "Using the forest to see the trees: A graphical model relating features, objects, and scenes," *Proc. Neural Info. Proc. System*, vol. 16, pp. 1499-1506, 2003.
3. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.
4. E. Gyftodimos and P. A. Flach, "Hierarchical Bayesian networks: A probabilistic reasoning model for structured domains," *Proc. European Conf. on Machine Learning*, pp. 25-36, 2003.

5. T. M. Strat and M. A. Fischler, “Context-based vision: Recognizing objects using information from both 2-D and 3-D imagery,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 10, pp. 1050-1065, 1991.
6. M. Marengoni, *et al.,* “Decision making and uncertainty management in a 3D reconstruction system,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol 25., no. 7, pp. 852-858, 2003.
7. A. Torralba, *et al.*, “Context-based vision system for place and object recognition,”*Proc. Intl. Conf. on Computer Vision*, pp. 273-280, 2003.
8. K. B. Korb and A. E. Nicholson, *Bayesian Artificial Intelligence*, CRC Press, 2003.
9. K. S. Hwang, *et al.*, “Bayesian network design for high-level context reasoning in uncertain indoor environment,”*Soft Computing Lab. Tech. Report*, 2005.
10. H. Reichenbach, *The Direction of Time*, University of California Press, 1956.

Design of a Simultaneous Mobile Robot Localization and Spatial Context Recognition System

Seungdo Jeong[1], Jonglyul Chung[1], Sanghoon Lee[1],
Il Hong Suh[2], and Byunguk Choi[2]

[1] Department of Electrical and Computer Engineering, Hanyang University
17 Haengdang-dong, Sungdong-gu, Seoul, 133-791 Korea
{kain,bellaw}@mlab.hanyang.ac.kr shlee@incorl.hanyang.ac.kr
[2] School of Information and Communications, Hanyang University
17 Haengdang-dong, Sungdong-gu, Seoul, 133-791 Korea
{ihsuh,buchoi}@hanyang.ac.kr

Abstract. In this work, we propose a simultaneous mobile robot localization and spatial context recognition system. The Harris corner detector and pyramid Lucas-Kanade optical flow are combined for robot localization. And, SIFT keypoints and its descriptors for the model-based object recognition and stereo vision technique are applied to spatial context recognition. The effectiveness of our proposed method is verified by experiments.

1 Introduction

While navigating in an environment to complete a delivery task, a mobile robot has to be able to recognize where it is, what the main objects in the scene are, and how main objects are spatially organized in the environment. Traditionally, place recognition, object recognition, and findings of spatial relations are considered separate problems. SLAM(Simultaneous Localization And Mapping) can be regarded as a popular technique to localize the mobile robot accurately and, simultaneously, to build a map of environment. To achieve SLAM, there are several different types of sensor modalities including laser range finder, sonar and monocular and/or stereo vision. A laser scanner is active, accurate, but slow and expensive, whereas sonar is fast and cheap, but usually very crude. Vision systems are passive and of high resolution.

Monocular vision-based SLAM technology as in [1] is highly desirable for a wide range of applications. However, the two dimensional vision-based SLAM technology in [1] cannot deal with object recognition and cannot find spatial relation of the objects.

Three dimensional vision approaches have been proposed in [2, 3]. Harris three-dimensional vision system DROID uses the visual motion of image corner features for 3D reconstruction [2]. Kalman filters are used for tracking features, and from the locations of the tracked image features, DROID determines both

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 945–952, 2005.

the camera motion and the 3D positions of the features. Ego-motion determination by matching image features is generally very accurate in the short to medium term. However, in a long image sequence, long-term drift can occur as no map is created.

In [3], Se describes a vision-based mobile robot localization and mapping algorithm, which uses Scale Invariant Feature Transformation(SIFT) keypoints as scale-invariant image features and natural landmarks in unmodified environments. The invariance of these features to image translation, scaling and rotation makes them suitable landmarks for mobile robot localization and map building. Feature viewpoint variation and occlusion are taken into account by maintaining a view direction for each landmark.

However, the SIFT stereo approach suffers from computational complexity, which can require 2-3 seconds of processing time per SIFT stereo image on contemporary PC hardware. With a robot velocity of 300mm/sec, this could result in a separation of up to 1m distance between consecutive SIFT stereo image locations. In this case, SIFT keypoints of the previous and current locations may be different, and exact point-to-point correspondence matching may not be possible. This would incur a 50% or greater error in relative position accuracy. In addition to the above difficulty, all these vision approaches are not concerned with recognition of objects and their spatial relations.

It is remarked that spatial relation is a key contextual information. To understand a local context, it is essential to firstly recognize objects within the local environment, and then to find out their spatial relations.

In this work, SIFT keypoints are used for model-based object recognition. And, object model includes three dimensional shape information. Once objects are detected, then Harris corner stereos are used to estimate spatial relations. In contrast to SIFT stereos, Harris corners can be calculated in a few hundred milliseconds on contemporary PCs, which can support 3D object motion tracking in a real-time. To ensure correct corner-to-corner correspondence between stereo images taken at consecutive locations, Lucas-Kanade optical flows are applied. This technique makes Harris corner detection less sensitive to changes in scale. Ego-motion of the mobile robot is then estimated by least square minimization of Harris corners of the matched landmark objects. After all, 3D relation between the mobile robot and main objects can be found, and thus 3D spatial relations between objects with respect to the robot can be also obtained. As a result, a mobile robot is able to recognize where it is, what the main objects in the scene are, how main objects are spatially organized in the indoor environment of the real world. Therefore, a mobile robot is able to complete the delivery task.

2 Overview of Simultaneous-Localization-and-Spatial-Context-Recognition System

Consider block diagram of our proposed simultaneous localization and spatial context understanding system as shown in Fig. 1. In Fig. 1, proposed system is composed of two functional parts. The robot localization module is designed to work with a sampling period of hundreds milliseconds. First of all, feature points are extracted by corner detection module within the images taken from the stereo camera. Next, the feature tracking module tracks the extracted feature points. These tracked feature points are used as point-based landmarks for the robot localization. The landmarks which is used for the robot localization are represented as 3D coordinates by using stereo matching. The robot localization module estimates the robot location by using the relationships between 3D landmarks and corresponding points projected to the current image.

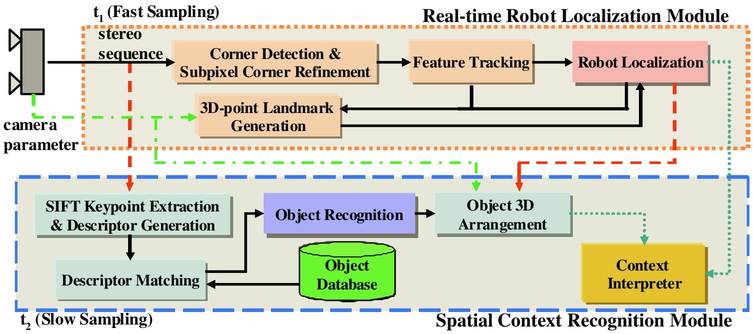


Fig. 1. Block Diagram of Simultaneous Localization and Spatial Context Understanding System

To understand the spatial context, the object recognition is essential. Our spatial context understanding module as in Fig. 1 is composed of the feature extraction module, the object recognition module, and the interpreter for spatial arrangement of the recognized objects. The context understanding module does not require real-time processing because there is no difficulty to understand the arrangement of the main objects and the spatial structure of local environment, even though this module does not operate at each frame. Therefore, we organize to work with slow sampling period contrary to the robot localization module, and to recognize the main objects exactly.

3 Design of a Real-Time Robot Localization Module

3.1 Detection of Harris Corners

The SIFT has been highlighted in the robot vision community recently, which is a method to extract and describe the feature point. SIFT uses Difference-of-Gaussian(DoG). Maxima and minima of the DoG images are detected by comparing a pixel to its 26 neighbors in 3×3 regions at the current and adjacent scales. These maxima and minima become the keypoints. In addition, the Gaussian image is down-sampled by a factor of 2, and the process is repeated. Therefore, the SIFT features are invariant to scale. A keypoint descriptor is created by first computing the gradient magnitude and orientation at each image sample point in the region around the keypoint location. To achieve orientation invariance, the coordinates of the descriptor and the gradient orientations are rotated relative to the keypoint orientation. Thus, SIFT is invariant to image rotation and scale, and robust across a substantial range of affine distortion, addition of noise, and change in illumination [4].

However, the algorithm to estimate the exact position of SIFT keypoint requires heavy computational cost. And, the position of a keypoint is not consistent. Thus, SIFT may not be a good feature candidate for a real time localization.

Alternatively, in this work, we will use Harris corner detector to extract the feature points in a real-time. The Harris corner detector [5, 6] is based on the autocorrelation function. To be specific, let $A(x, y)$ be defined by

$$A(x,y) = \begin{bmatrix} \sum (I_x(x_k, y_k))^2 & \sum I_x(x_k, y_k) I_y(x_k, y_k) \\ \sum I_x(x_k, y_k) I_y(x_k, y_k) & \sum (I_x(x_k, y_k))^2 \end{bmatrix}, \tag{1}$$

where (x_k, y_k) are the points in an window centered on (x, y). Then, corner points are detected if the autocorrelation matrix A has two significant eigenvalues.

In order to verify the Harris corner is more suitable for real time localization than SIFT, Table 1 shows the performance comparison of the Harris corner detector and the SIFT algorithm. The Harris corner detector includes the corner extraction and the sub-pixel refinement, whereas, the SIFT algorithm includes the extraction of keypoints and the generation of descriptors. Although, the number of feature points of the Harris corner detector and the SIFT algorithm looks similar, the Harris corner detector has relatively low computational cost when compared to the SIFT algorithm. Thus, Harris corner detector is more suitable for the real-time processing than SIFT. In Table 1, the unit of computation time is millisecond.

Table 1. Feature extraction time per image

Image size	Harris corner detector				SIFT algorithm			
	Feature Points	Max	Min	Average	Feature Points	Max	Min	Average
176 × 144	252.8	45	14	27.65	270.4	860	406	651.24
320 × 240	474.1	47	14	28.86	492.1	3157	1875	2292.57
640 × 480	696.4	125	31	62.35	715.7	8891	6469	7660.8

3.2 Tracking of Corner Features

To track the camera motion, it is necessary to obtain the feature points in the current frame corresponding to the feature points in the previous frame. Note that the corner points satisfy the local smoothness constraint. In other words, the motion of neighborhood points about a corner is almost same. Supposing the consecutive frames called H and I, the brightness of the corresponding point is the same. That is, the motion of a pixel is represented as

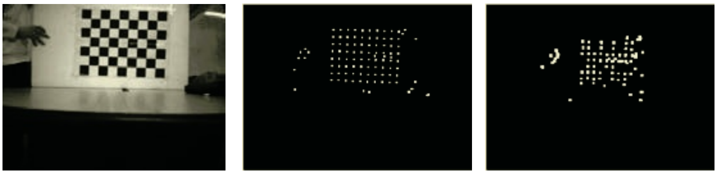
$$H(x, y) = I(x + u, y + v). \tag{2}$$

With the first-order Taylor expansion of $I(x + u, y + v)$, we can derive the final optical flow equation as

$$I_t + \nabla I \cdot [u\ v] = 0. \tag{3}$$

If we estimate the motion of one pixel only, there will be the aperture problem. To resolve this aperture problem, we can apply the local smoothness constraint; the motions of neighborhood pixels are almost same. Therefore, the matrix form of the optical flow equation should be considered to include $n \times n$ window around the feature point.

The Lucas-Kanade optical flow is suitable to the small motion of one pixel difference [7]. However, we can hardly expect such a small motion as in the practice. Thus, we apply the pyramid Lucas-Kanade optical flow. The pyramid Lucas-Kanade optical flow first requires to generate the Gaussian image pyramid. Then optical flow is estimated from the lowest level of sub-sampled image to original image iteratively by using image warping and up-sampling process.



(a) Input Image (b) Harris Corner (c) SIFT

Fig. 2. Comparison of the Harris corner and the SIFT stereo

The experimental result for stereo matching using the Harris corner detector following the pyramid Lucas-Kanade optical flow and using the SIFT algorithm is shown in Fig. 2. Our algorithm not only extracts the feature points with coherency, but also tracks the feature points exactly as shown in Fig. 2. However, the SIFT algorithm has little coherency though the extracted feature points are invariant to the variation of image. This shows that there exists a difference in the repeatability in the sense that the same feature point is extracted consecutively.

3.3 Robot Localization

The feature points extracted from the initial frame are used as 3D point landmark. Note that we use the stereo camera as input sensor, we obtain stereo image pair. The 3D coordinate of the extracted feature point is calculated by using disparity of that in stereo camera. The stereo camera is calibrated previously. When we know the relationship exactly between 3D coordinates of the landmark and corresponding 2D coordinates which are projected to image, the projection matrix indicates the mapping relationship between points of 3D space and points of image. The projection matrix is given by

$$\begin{bmatrix} x \\ y \\ w \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} R \,|t \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{4}$$

where R is 3×3 rotation matrix and t is 3×1 translation matrix.

Projection matrix is composed of two parts. One is the intrinsic parameter matrix which includes internal information of camera. This matrix represents the relationship between camera coordinate system and image coordinate system. In intrinsic parameter matrix, f_x, f_y represent the focal length for each direction respectively, (c_x, c_y) represents the principal point. Another part of the projection matrix is the extrinsic parameter matrix. This matrix represents the translational and rotational relationship between the 3D space coordinate system and the camera coordinate system [8]. With that the intrinsic parameter matrix is already obtained by in processing of the stereo camera calibration. Therefore, if the projection matrix is estimated with coordinates of corresponding points in 3D space and in 2D image, we can obtain the motion of camera.

It is the non-linear estimation problem to estimate the projection matrix with numerous corresponding points. In this work, we estimate rotation and translation components by using Levenberg Marquardt least square minimization method [9]. The robot localization is accomplished by converting camera motion to the 3D space coordinate system.

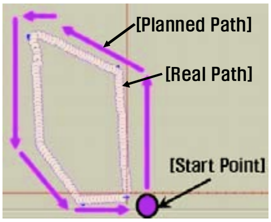


Fig. 3. Robot localization result

In order to verify the localization performance for the successful delivery task, we observe how precise the robot recognizes location and it returns to the start position exactly with the planed path. Fig. 3 shows the localization result by using our localization module for the planned path. The robot moves from start position and return to start position navigating about 6.5 meter including with moving in front and rotating. Last position error is about 10 centimeter by using the proposed localization method. Each error of moving robot position is shown in the Table 2. Each numerical value is surveyed per 50 millimeters.

4 Design of Spatial Context Recognition Module

4.1 Object Recognition

It is required to extract the precise feature points more than the Harris corner for object recognition. In this work, the SIFT keypoints and their descriptors are used for model-based object recognition [10]. The object database is previously builded in such a way that this database includes the information of the

Table 2. Error of the Robot Localization

	Harris Corner + Pyramid L-K		
	Maximum	Minimum	Average
X(millimeter)	17.72	0.52	4.896
Y(millimeter)	5.42	0.14	3.12
Z(millimeter)	18.22	1.62	6.146
θ(degree)	0.418	0.021	0.071

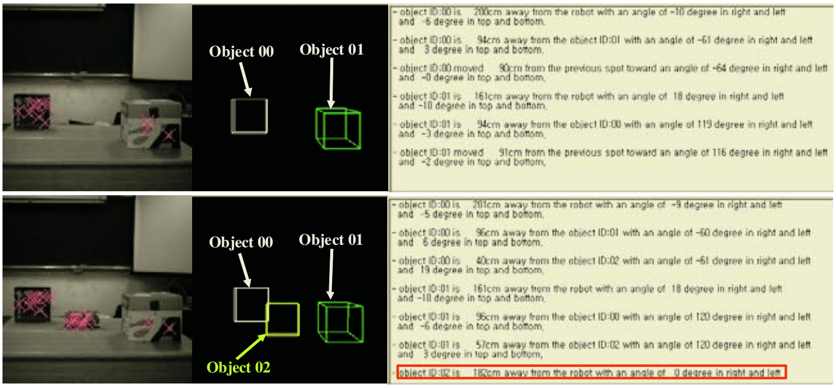
SIFT descriptor for each image patch. The object recognition is accomplished by matching between the SIFT keypoints extracted from current scene with the SIFT keypoints of each object in the object database.

4.2 Extraction of Spatial Relations of Objects

The recognized object has only 2D coordinate in current scene yet. So, to represent relative arrangement of each object in the 3D space we need to know the 3D coordinate of each object. The disparities of the SIFT keypoints within each object are calculated individually and then, the 3D coordinate of each keypoint is obtained by using these disparities. Because the 3D coordinate of each keypoint is already classified into each object, the region including all 3D keypoints of object is the 3D location of the object. However, this location refers to the current camera. So, we need to know the location of the current camera in order to obtain the position of the object in the absolute space. The current camera position can be acquired from the robot localization module in the real-time processing sub system. Therefore, we can finally represent the 3D coordinate of respective object with respect to the absolute coordinate system using the robot localization information and 3D location of each object with respect to the current camera. Locations of objects interpreted by interpreter module are informed to robot localization module to correctly compensate the cumulated error caused by long-term operations of the Harris corner detector and pyramid Lucas-Kanade optical flow. Fig. 4 shows an example of recognition of spatial relations of objects. This example shows that the interpreter module classifies new object and exactly interprets its spatial relation with respect to previous objects, whenever it appears newly.

5 Conclusion

In this work, we have proposed a method to not only localize a mobile robot but also recognize the spatial context. This was implemented by effective hybridization of several paradigms; for localization, Harris corner detector and pyramid Lucas-Kanade optical flow. And for recognition of spatial relations, 2D and 3D SIFT keypoints were employed. Our proposed method was successfully applied to a mobile robot navigation.



```
- object ID:00 is    200cm away from the robot with an angle of -10 degree in right and left
  and  -6 degree in top and bottom.
- object ID:00 is     94cm away from the object ID:01 with an angle of -61 degree in right and left
  and  3 degree in top and bottom.
- object ID:00 moved     90cm from the previous spot toward an angle of -64 degree in right and left
  and  -0 degree in top and bottom.
- object ID:01 is    161cm away from the robot with an angle of  18 degree in right and left
  and -10 degree in top and bottom.
- object ID:01 is     94cm away from the object ID:00 with an angle of 119 degree in right and left
  and  -3 degree in top and bottom.
- object ID:01 moved     91cm from the previous spot toward an angle of 116 degree in right and left
  and  -2 degree in top and bottom.
```

```
- object ID:00 is    201cm away from the robot with an angle of  -9 degree in right and left
  and  -5 degree in top and bottom.
- object ID:00 is     96cm away from the object ID:01 with an angle of -60 degree in right and left
  and  6 degree in top and bottom.
- object ID:00 is     40cm away from the object ID:02 with an angle of -61 degree in right and left
  and  19 degree in top and bottom.
- object ID:01 is    161cm away from the robot with an angle of  18 degree in right and left
  and -10 degree in top and bottom.
- object ID:01 is     96cm away from the object ID:00 with an angle of 120 degree in right and left
  and  -6 degree in top and bottom.
- object ID:01 is     57cm away from the object ID:02 with an angle of 120 degree in right and left
  and  3 degree in top and bottom.
- object ID:02 is    182cm away from the robot with an angle of   0 degree in right and left
```

Fig. 4. Spatial relations of objects

Acknowledgement

This work is supported by the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Korea Ministry of Commerce, Industry and Energy.

References

1. A. J. Davison: Real-Time Simultaneous Localization and Mapping with a Single Camera. Proceedings of International Conference on Computer Vision. (2003) 1403–1411
2. C. Harris: Geometry from visual motion. Active Vision. MIT Press. (1992) 264–284
3. S. Se, D. Lowe, and J. Little: Mobile Robot Localization and Mapping with Uncertainty using Scale-Invariant Visual Landmarks. International Journal of Robotics Research. (2002) 735–758
4. D. Lowe: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision. Vol. 60 (2004) 91–110
5. C. Harris and M. J. Stephens: A combined corner and edge detector. Alvey Vision Conference. (1988) 147–152
6. C. Schmid, R. Mohr, and C. Bauckhage: Evaluation of Interest Point Detectors. International Journal of Computer Vision. Vol. 37, No. 2 (2000) 151–172
7. B. Lucas and T. Kanade: An interative image registration technique with an application to stereo vision. Proc. DARPA IU Workshop. (1981) 121–130
8. R. Hartley and A. Zisserman: Multiple View Geometry in Computer Vision. Cambridge University Press. (2000)
9. W. Press, S. Teukolsky, W. Vetterling, and B. Flannery: Numerical Recipes in C++. Cambridge University Press. (2002)
10. S. Helmer and D. Lowe: Object Class Recognition with Many Local Features. Workshop on Generative Model Based Vision. (2004)

Mining Temporal Data: A Coal-Fired Boiler Case Study

Andrew Kusiak and Alex Burns

Intelligent Systems Laboratory, Industrial Engineering
3131 Seamans Center, The University of Iowa
Iowa City, IA 52242 – 1527, USA
andrew-kusiak@uiowa.edu

Abstract. This paper presents an approach to control pluggage of a coal-fired boiler. The proposed approach involves statistics, data partitioning, parameter reduction, and data mining. The proposed approach was tested on a 750 MW commercial coal-fired boiler affected with a fouling problem that leads to boiler pluggage that causes unscheduled shutdowns. The rare-event detection approach presented in the paper identified several critical time-based data segments that are indicative of the ash pluggage.

1 Introduction

The ability to predict and avoid rare events in time series data is a challenge that could be addressed by data mining approaches. Difficulties arise from the fact that often a significant volume of data describes normal conditions and only a small amount of data may be available for rare events. This problem is further exacerbated by the fact that traditional data mining does not account for the time dependency of the temporal data. The approach presented in this paper overcomes these concerns by defining time windows.

The approach presented in this paper is based on the two main concepts. The first is that the decision-tree data-mining algorithm captures the subtle parameter relationships that cause the rare event to occur [1]. The second concept is that partitioning the data using time windows provides the ability to capture and describe sequences of events that may cause the rare failure.

2 Event Detection Procedure

In the case study discussed in the next section rare events can be detected by applying the five step procedure. These five steps include:

Step 1: Parameter Categorization

The parameter list is divided into two categories, response parameters and impact parameters. Response parameters are those that change values due to a rare event or a failure, e.g., an air leak in a pressurized chamber.

Impact parameters are defined as parameters that are either directly or indirectly controllable and may cause the rare event. These are the parameters that are of greatest interest for the determination of rare events.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 953–958, 2005.

Step 2: Time Segmentation

Time segmentation deals with partitioning and labeling the data into time windows (TWs). A time widow is defined as a set of observations in chronological order that describe a specified amount of continuous observations. This step allows the data mining algorithms to account for the temporal nature of the data. The most effective method to segment the data is by determining/estimating the approximate date of failure and set that as the last observation of the final time window.

Step 3: Statistical and Visual Analysis

This step involves statistical analysis of the data in each time period that was designated in the previous step. Process shifts, changes in variation, and mean shifts in parameters are helpful in indicating that the appropriate time windows and parameters were selected.

Step 4: Knowledge Extraction

Data mining algorithms discover relationships among parameters and an outcome in the form of IF … THEN rules and other constructs (e.g., decision tables) [1], [5]. Data mining is natural extension of more traditional tools such as neural networks, multivariable algorithms, or traditional statistics. In the detection of rare events, the decision-tree and rule-induction algorithms are explored for two significant reasons. First, the algorithms generate explicit knowledge in the form understandable by a user. The user is able to understand the extracted knowledge, assess its usefulness, and learn new and interesting concepts. Secondly, the data mining algorithms have been shown to produce highly accurate knowledge in many domains.

Step 5: Analysis of Knowledge and Validation

This step deals with validation of the knowledge generated by the data mining algorithm. If a validation data set is available it should be used to validate the accuracy of the rules. If no similar data is available then unused data from the analysis or a 10-fold cross-validation can be utilized [6].

3 Power Boiler Case Study

The approach proposed in this research was applied to power plant data. Data mining algorithms are well suited for electric power applications that produce hundreds of data points at any time instance.

This case study deals with an ash fouling condition that causes boiler shutdowns several times a year on a commercial 750 MW tangentially-fired coal boiler. The ash fouling causes a build up of material and pluggage in the reheater section of the boiler. Once the build up becomes substantial the boiler performance is negatively affected. This leads to the derating and the eventual shutdown of the boiler. The cleaning of the boiler during the shutdown requires 1 to 3 days. This problem is made more difficult by the fact there is no method to determine the level of ash build up without shutting down the boiler to physically inspect the area. Furthermore, in analysis all parameters were within specifications, so there was no obvious single parameter that is causing the pluggage.

To investigate the problem considered in this paper, data was collected on 173 different boiler parameters. This included flows, pressures, temperatures, controls, demands, and so on. The data was collected in one-minute intervals over the course of three months. The data collection began directly following a shutdown where the reheater section of the boiler had no pluggage. The collection period ended approximately three months later when the boiler had to be shutdown for pluggage removal. This data set contained over 168,000 observations.

The list of 173 parameters, which included both response and impact parameters, was analyzed. The list was reduced to include twenty-six impact parameters. This parameter categorization and reduction was accomplished with the assistance of domain experts as well as statistical analysis such as correlation and multivariate analysis.

The initial step for time segmenting the data was to determine an approximate date for the failure event. In this application the failure event was defined by the date when the boiler was derated due to the pluggage. The cause of the shutdown was confirmed through visual inspection of the affected region. This date was then set to be the last day of the final time window (TW6).

The windows were set to be approximately one week long. A week was chosen for several reasons. First, the boiler was inspected approximately one month prior to its derating. During the inspection the reheater section of the boiler was completely free of ash. This information provided the knowledge that the pluggage required less the one month to manifest itself to the point of shutdown. It was hypothesized that the pluggage requires several days to build up. Based on this information one week was deemed to be an adequate time window. One week also provided a sufficient number of observations (over 10,000 per window) for the data mining algorithms.

Using the derate date and a one-week-long time window, the data was divided into six time windows shown in Figure 1. Time window 1 (TW1) was included to ensure that there was adequate data to describe normal operating conditions.

There appears be a process shift between time windows 3 and 5 in Figures 1. The west tilt demonstrates a mean shift during window three and the hot reheat steam temperature displays a mean shift as well as a large increase in variation starting in time window four and culminating in window five. The results of this analysis lead to the hypothesis that the events that lead to the eventual pluggage occur between time windows three and five. It also confirms the selection of parameters and window size.

The data mining approach was then applied to the data set to predict the predefined time windows (decision parameter). The algorithm produced a set of rules that described the parameter relationships in each time window.

The knowledge extracted by the algorithm had an overall 10-fold classification accuracy of 99.7%. The confusion matrix (absolute classification accuracy matrix) is shown in Figure 2. The matrix displays the actual values and the values predicted by the rules during the cross-validation process.

It can be seen from the data in Figure 2 that there are few predicted values that are off by more than one time window from the actual window. The results provided in the confusion matrix provide a high confidence in the proposed solution approach.

Another test data set was extracted from the week following time window 1 and was labeled time window 2 (Test TW2). The last portion of the data (Test TW3) was obtained from the week after the generator was derated and the outcome was labeled time window 6 (TW6). The total test set contained over 30,000 observations.

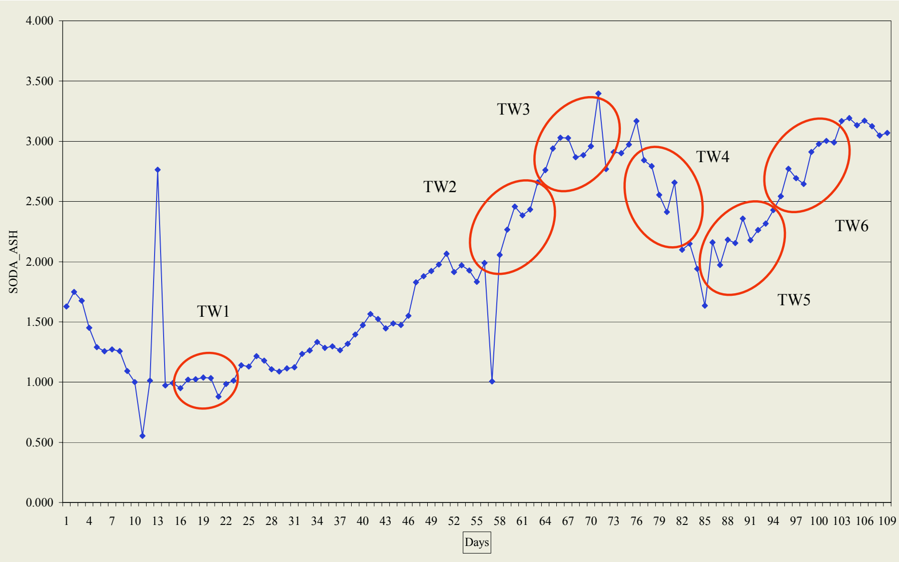


Fig. 1. Time windows for ash fouling application

		Predicted Value					
		TW1	TW2	TW3	TW4	TW5	TW6
Actual Value	TW1	11513	7				
	TW2	17	11497	5			1
	TW3	11		11483	9	5	12
	TW4			13	11461	14	15
	TW5			2	15	12913	30
	TW6	1		1	3	49	12906

Fig. 2. Confusion matrix

		Predicted Value		
		Normal	Fault	Accuracy
Actual Value	Normal	19630	370	98.15%
	Fault	2683	7305	73.14%

Fig. 3. Cross-validation results for the test data set

The rules and knowledge that were extracted from the original data set were then tested using the test data set. For purposes of analysis time windows 1 – 3 were considered normal and time window 4 – 6 were considered faulty. The resulting confusion matrix is shown in Figure 3.

The rules accurately predicted the normal cases, but they were not as effective in predicting the fault cases. This is most likely explained by the fact that the test data labeled, time window 6, was extracted after the boiler had been derated. The derating of the boiler significantly changes the combustion process and was not included in the original data set. In spite of this, the overall classification accuracy of the test data set is greater than 89%. The high cross-validation accuracy indicates that the rules accurately capture the changes in the process that lead to the ash fouling, pluggage, derating, and eventual shutdown of the boiler.

4 Future Research

Event detection for control advisory systems has also been successfully demonstrated for applications that are dynamic and involve rare and catastrophic events [4]. Finch *et al.* [2] developed expert diagnostic information system, MIDAS, to alert users to abnormal transient conditions in chemical, refinery, and utility systems [3].

The approach presented in this research produced rule sets that can be utilized for the development of a meta-control system. Integrating concepts from expert advisory systems and intelligent power control systems will form the meta-control system architecture for the avoidance of the ash pluggage.

5 Conclusion

In this paper a data mining approach to predict failures was proposed and successfully implemented. The research utilized parameter categorization and time segmentation to overcome the limitation of traditional data mining approaches applied to temporal data. The proposed approach produced a knowledge base (rule set) that accurately described the subtle process shifts and parameter relationships that eventually may lead to the detection and avoidance of failures.

This approach was applied to a commercial tangentially-fired coal-boiler to detect and avoid an ash fouling pluggage that eventually leads to boiler shutdown. The approach produced a rule set that was over 99.7% accurate. The knowledge base was also validated with a separate test data set that has predicted failures with accuracy of over 89.8%.

The discovered knowledge will be used to develop an advance warning system reducing the number of boiler shutdowns. The intelligent warning system will have a significant economic impact. This translates into reduced cost to the consumer and a more efficient power industry.

References

1. Quinlan, J.R., “Induction of decision trees,” *Machine Learning,* vol. 1, no. 1, pp. 81-106, 1986.
2. Branagan, L. A., and Wasserman, P. D., "Introductory use of probabilistic neural networks for spike detection from an on-line vibration diagnostic system", *Intelligent Engineering Systems Through Artificial Neural Networks*, vol. 2, pp. 719-724, 1992.

3. Finch, F. E., Oyeleye, O. O., and Kramer, M. A., "Robust event-oriented methodology for diagnosis of dynamic process systems", *Computers & Chemical Engineering*, vol.14, no. 12, pp. 1379-1396, Dec, 1990.
4. Pomeroy, B. D., Spang, H. A., and Dausch, M.E., "Event-based architecture for diagnosis in control advisory systems", *Artificial Intelligence in Engineering*, vol. 5, no. 4, pp. 174-181, Oct, 1990.
5. Pawlak, Z., Rough Sets: Theoretical Aspects of Reasoning About Data, Boston: Kluwer, 1991.
6. Stone, M. "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society*, vol. 36, pp.111-147, 1974.

Mining Classification Rules Using Evolutionary Multi-objective Algorithms

Kalyanaraman Kaesava Kshetrapalapuram and Michael Kirley

Department of Computer Science and Software Engineering
University of Melbourne, Australia
{kkkshe,mkirley}@cs.mu.oz.au

Abstract. Evolutionary-based methods provide a framework for mining classification rules, that is, rules that can be used to discriminate between data organized in several classes. In this paper, we propose a novel multi-objective extension for the standard Pittsburg approach. Key features of our model include (a) variable length chromosomes, implemented using an active bit string (mask), and (b) fitness evaluation and selection based on restricted non-dominated tournaments. Extensive numerical simulations show that the proposed algorithm is competitive with – and indeed outperforms in some cases – other well-known machine learning tools using benchmark datasets.

1 Introduction

The benefits of finding trends in large volumes of data have driven the development of data mining for over a decade. The most common data mining procedure, classification, refers to the induction of rules that discriminate between data organized in several classes so as to gain predictive power [5]. Inducing such rules essentially involves the simultaneous optimization of several objectives (eg. accuracy, coverage and completeness) over a training data set. Classification is thus an example of a multi-objective optimization problem.

Evolutionary techniques (eg. genetic algorithms, evolutionary strategies) have become the *de facto* standard for multi-objective search and optimization problems [3, 9]. There are some instances of evolutionary multi-objective applications for classification rule induction in the literature (eg. [1, 5, 8]), however, there has only been limited use. This is somewhat surprising given that genetic algorithms deal with attribute interaction very effectively because they explore the entire (global) search space simultaneously [5].

In this paper, we propose a new data mining model that evolves a range of compromise or trade-off solutions for classification tasks. Here, individuals encode a full and variable-length rule set solution, which is evaluated in terms of both accuracy and coverage. The efficacy of the proposed model is reported using benchmark datasets.

2 Background

2.1 Evolutionary Rule Induction Techniques

The mining of classification rules typically involves a two step process. Firstly, the data under investigation is divided into mutually exclusive and exhaustive

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 959–965, 2005.

sets, referred to as the training and the test sets. The goal of the data mining algorithm is to discover rules by accessing the training set only. Once the training has finished, the predictive performance of the discovered rules is evaluated using the test set, which was not seen during training.

Genetics-based machine learning systems, such as classifier systems (and variants thereof) have been used to tackle this particular data mining task. Evolutionary–based models can be divided into two different approaches. In the *Michigan* model, each individual encodes a single prediction rule. Here, the objective is to discover a set of rules, rather than a single rule. In contrast, in the *Pittsburgh* model each individual encodes a set of prediction rules [4]. That is, a single individual concatenates a population of classifiers, which is then subjected to the usual evolutionary operators [4, 7]. In this study, we employ a version of the Pittsburgh model because we are interested in evaluating the quality of the rule set as a whole (that is, the best concatenated rule) rather than the quality of a single rule.

2.2 Evolutionary Multi-objective Algorithms

In many real-world search and optimization tasks, we are often confronted with a problem involving several incommensurable and often conflicting objectives. For example, in classification task there may be a trade-off between classifier accuracy and coverage. For such problems, a family of equivalent *non-dominated* compromises — the Pareto-optimal set — represent solutions for the given problem [3]. These solutions are optimal in the wider sense that no other solution in the search space is superior to them when all objectives are considered. Evolutionary algorithms are particularly suitable for solving multi-objective optimization problem because they are not sensitive to different Pareto front shapes and are able to find solutions located in non convex or discontinuous zones. In addition, they can effectively deal with stochastic characteristics, uncertainties or with noise within objective functions.

In order to maintain a spread of solutions across the entire Pareto Front, the multi-objective evolutionary algorithms should promote population diversity as well as incorporating elitism models. A large number of algorithms, including SPEA, NSGA 2 (see[3] for an overview), have been described in the literature, which have these characteristics. In this study, the algorithm we have implemented is a variant of NPGA [6], which makes use of restricted tournament selection techniques to foster niching in the population, with multi-objective fitness evaluation.

3 A Model for Evolutionary Multi-objective Classification

In this section, we describe a framework for classification rule induction using a multi-objective evolutionary algorithms. Firstly, the underlying genetic encoding is described. This is followed by a description of the multi-objective algorithm fitness evaluation and selection.

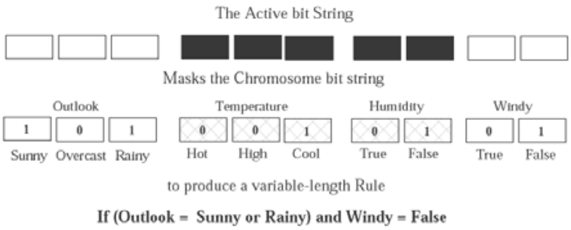

The Active bit String
Masks the Chromosome bit string
Outlook
Temperature
Humidity
Windy
1
0
1
0
0
1
0
1
0
1
Sunny
Overcast
Rainy
Hot
High
Cool
True
False
True
False
to produce a variable-length Rule
If (Outlook = Sunny or Rainy) and Windy = False

Fig. 1. Chromosome encoding

3.1 Representation and Operators

The chromosome encoding consists of two components (a) the actual rules, and (b) an *active bit* string. The example in Fig. 1 describes this implementation using a simple example (if $outlook$ = ($sunny$ or $rainy$) and $windy = false$). Note that *temperature* and *humidity* do not participate in the rule because they are masked by an active bit string that is also subject to evolution. The active bit string thus implements the idea of a variable-length chromosome. The class for the rule is not part of the encoding and does not evolve with the chromosome.

Each chromosome in the population is initialized by turning on exactly one attribute value randomly chosen. All genes are initially made active, so they can participate in the rule. Standard evolutionary operators are employed in our model. A randomly selected crossover operator – one-point, two-point or uniform – is applied to both the chromosome and the active bit string with equal probability. Mutation of the chromosome involves the random flipping of a bit, while mutation of the active bit string involves flipping an entire gene, each occurring with equal probability.

3.2 Fitness Evaluation and Selection

For a given classification task, two fitness objectives are used in our algorithm (a) *confidence* or accuracy – proportion of the rules satisfied by the antecedent that belong to the predicted class of the rule), and (b) *coverage* – proportion of tuples satisfied by the rule antecedent.

In the selection phase of the algorithm, we implement a version of *restricted tournament selection* (RTS). In the RTS algorithm, two chromosomes (say A and B) are chosen randomly from the current population. Meanwhile, two random (possibly overlapping) subpopulations (whose size is determined by the crowding factor cf), are also chosen, (say P_A and P_B). The crossover operator is applied to A and B and the resulting chromosomes (A' and B') are mutated to produce A'' and B'' respectively. The chromosome most similar to A'' in P_A (say A^*) and similarly B^* are found. A^* is then played off against A''. If A'' wins, then it replaces A^* in the population. Similarly, with B. The process is repeated n (population size) times. This whole procedure constitutes one generational change. The algorithm is shown in Fig. 2.

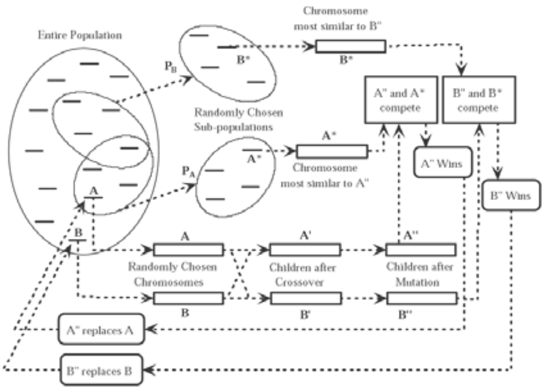


Fig. 2. Restricted tournament selection algorithm

If either competitor dominates the other, it wins. Otherwise, the one with the larger difference between the number of individuals that it dominates to the number of individuals that dominate it, wins. In the case of a tie, the winner is randomly chosen. Diversity is maintained because competition is only between similar chromosomes.

We assign a chromosome/rule to the class with the greatest number of training instances that are covered by the rule. It is therefore possible for the entire population to mine rules for a single class, making accurate classification impossible. To overcome this problem, we use two alternative variations where the population is divided into segments, each committed to finding rules for one class. The population is split either equally among the classes or proportionally, based on the fraction of training instances of that class. Such a segmentation is then equivalent to running the RTS algorithm as many times as there are classes, but in parallel.

The support of a rule is some combination of the rule's fitness vector (like product of the confidence and coverage). A testing instance is classified as the class with the most accumulative support of the rules from the final generation that apply to it. Various combinations of the rule's fitness vector were experimented with and we found that there was no combination that consistently produced the best accuracy.

4 Simulations and Results

To evaluate the effectiveness of our algorithm, we compare its performance against other well-known machine learning techniques using six benchmark data sets (see table 1) from the UCI ML Laboratory [2]. They have a large number of attributes and/or classes and have often been cited in related papers. All tests were performed using stratified ten-fold cross-validation.

Table 1. Datasets Used

Dataset	Instances	Attributes	Classes
Dermatology	366	34	6
Diabetes	768	8	2
Glass	214	10	7
Sonar	208	60	2
Vehicle	846	18	4
Promoters	106	57	2

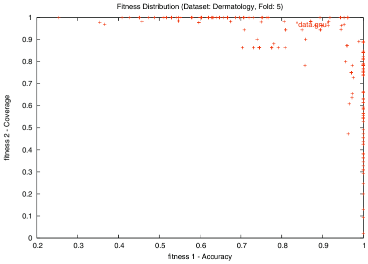


Fig. 3. Fitness distribution after 50 generations

4.1 Parameters

Data pre-processing was done using the WEKA data mining program developed by the University of Waikato (New Zealand). We applied the unsupervised Discretizer Filter with the option to discretize attributes with continuous values into equal-frequency bins turned on. Instances were internally stored as bit vectors and the corresponding bit of a missing attribute value was not set.

In our multi-objective algorithm, the rate of crossover is set to 0.85. The mutation rate was set to 0.15. The rate is higher than commonly used. This is because for chromosomes with no modification, they will either compete with themselves, get thrown away or enter as duplicates. For all experiments, we used 75 for the number of generations. For classification without population segmentation, we set $cf = 10\%$, consistent with the literature. The population size was set to 50 times the number of classes for fixed-segment segmentation, and otherwise fixed at 200.

4.2 Results

The plot in Fig. 3 illustrate the approximate Pareto-front after 50 generations for a typical trial. An inspection of the plot reveals that the RTS algorithm is able to direct the fitness vectors towards the top-right corner (both parameters - accuracy/confidence and coverage need to be maximized) of the graph while maintaining diversity across the Pareto Front. The individual non-dominated solutions found at the end of the training phase produce rule sets with high ac-

Table 2. Comparison of Error Rates

	Dataset	Our Model	One R	Naive Bayes	C4.5 Decision Tree
1	Dermatology	0.07	0.02	0.06	0.50
2	Diabetes	0.18	0.25	0.28	0.25
3	Glass	0.35	0.28	0.41	0.43
4	Promoters	0.21	0.10	0.19	0.30
5	Sonar	0.19	0.21	0.34	0.31
6	Vehicle	0.41	0.37	0.33	0.48

curacy levels when evaluated against the testing set, thus showing they generalize well.

Table 2 shows error rates of our algorithm (averaged over all runs) compared with those produced by WEKA's implementation of well-recognized algorithms on exactly the same datasets (after pre-processing). Our results are significantly better for the Diabetes and Sonar data sets ($p<0.05$). For the other data sets, our model is competitive and in some instances outperforms individual techniques for a given data set.

5 Discussion and Conclusion

In a classification task, the goal is to use previously observed data to construct a model, which is able predict the categorical or nominal value (the class) of a dependent variable given the value of the independent variables. In this context, we want the discovered model to have a high predicative accuracy. In addition, we would also like the evolved rule sets to be comprehensible and to have high coverage.

In this study, we have cast the problem of classification rule induction in terms of a multi-objective search and optimization problem. The model described here, is capable of deriving classification rules with varying degrees of generality. Key features of the model include (a) variable length chromosomes, implemented using an active bit string (mask), and (b) fitness evaluation and selection based on restricted non-dominated tournaments. Standard Pittsburg models have a tendency to generate rules with satisfactory accuracy levels, but relatively small coverage for a given dataset. The restricted non-dominated selection techniques used in our model go some of the way to improving the overall performance of the classifier. We have investigated alternative techniques for segmenting the population, calculating rule supports, and other evolutionary parameters. Extensive evaluation of our model was conducted using stratified ten-fold cross-validation. It is interesting to note, that although our algorithm does not outperform the other techniques when the a dataset has many classes, it is competitive. Even in these cases, it is never the worst performer. It out performs the C4.5 decision tree algorithm in every dataset. We believe the reason might be that the datasets we chose are particularly complex and involve a fair amount of attribute interaction. Since decision trees make decisions on single attributes, they are unable to cope with attribute interaction very well.

Data mining and knowledge discovery differ from traditional machine learning approaches in that the focus is typically on discovering interesting, novel or surprising rules. We are currently experimenting with extending our algorithm to focus on association rule induction. In future work, we will also examine alternative techniques to control the functioning of the active bit string. For example, our work can be extended to mine association rules using concepts such as gene regulation to model complex attribute interaction.

References

1. Bedingfield, S.E and Smith, K.A. : Evolutionary Rule Generation Classification and Its Application to Multi-class data. In ICCS 2003. Lecture Notes in Computer Science. 2660. pp868-876. (2003)
2. Blake C.L., Merz, C.J.: UCI Repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences (1998)
3. Coello Coello, C.A., Van Veldhuizen, D.A. and Lamont, G.B. : EA for Solving Multi-Objective Problems. Kluwer Academic Publishers. (2002).
4. DeJong, K. A., Spears, W. M., and Gordon, D. F. : Using genetic algorithms for concept learning. Machine Learning, 13 (2/3), (1993) 161–188
5. Freitas, A.A.: On Objective Measures of Rule Surprisingness. In: Principles of Data Mining and Knowledge Discovery (1998) 1–9
6. Harik, G.R.: Finding Multimodal Solutions Using Restricted Tournament Selection. In: Larry Eshelman (ed.): Proc. of 6thh Intl. Conf. on GAs. Morgan Kaufmann, San Francisco, CA (1995) 24–31,
7. Llora, X., Garrell, J.M.: Co-evolving Different Knowledge Representations with fine-grained Parallel Learning Classifier Systems. In: Proc. Genetic and Evol. Comp. Conf. (GECCO2002). Morgan Kaufmann (2002)
8. Llora, X., Goldberg, D.E., Traus, I., Bernado, E.: Accuracy, Parsimony, and Generality in Evolutionary Learning Systems via Multiobjective Selection. In: Proc. 5th Intl. Workshop on Learning Classifier Systems (2002)
9. Zitzler, E., Deb, K., Thiele, L.: Comparison of Multiobjective Evolutionary Algorithms: Empirical Results. In: Evol. Comp., Num. 2, Vol. 8 (2000) 173–195

On Pruning and Tuning Rules for Associative Classifiers*

Osmar R. Zaïane and Maria-Luiza Antonie

University of Alberta, Edmonton, Canada
{zaiane,luiza}@cs.ualberta.ca

Abstract. The integration of supervised classification and association rules for building classification models is not new. One major advantage is that models are human readable and can be edited. However, it is common knowledge that association rule mining typically yields a sheer number of rules defeating the purpose of a human readable model. Pruning unnecessary rules without jeopardizing the classification accuracy is paramount but very challenging. In this paper we study strategies for classification rule pruning in the case of associative classifiers.

1 Associative Classifiers and Their Massive Model

Association rules are typically known as an important and common means for market basket analysis. However, it has been observed that association rules could be used to model relationships between class labels and features from a training set [4]. Therefore, association rules were used to efficiently build a classification model from very large training datasets. Since then, many associative classifiers were proposed mainly differing in the strategies used to select rules for classification and in the heuristics used for pruning rules [6, 7, 9]. Among the many advantages of associative classifiers we can highlight four major ones:

- The training is very efficient regardless of the size of the training set;
- Training sets with high dimensionality can be handled with ease and no assumptions are made on dependence or independence of attributes;
- The classification is very fast;
- The classification model is a set of rules easily understandable by humans and can be edited.

The problems with associative classifiers are also remarkable. First, they inherit two complicated parameters from association rule mining, namely *support* and *confidence*. These are difficult to set and tune. Second, association rule mining generates a sheer number of rules commonly outnumbering the observations in the training set. This defeats the purpose of readability of the classification model since no human would be willing to sift through hundreds of thousands of rules for editing purposes. This leads to two other issues: How can we reduce

* Partially supported by Alberta Ingenuity Fund, iCORE and NSERC Canada

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 966–973, 2005.

the number of rules in the model and how can we effectively select rules to apply during classification? In this paper we address one of these issues: the reduction of classification rules. This problem is challenging because the goal is to prune rules while preventing the accuracy of the classifier from dipping.

1.1 Motivation and Contributions

Our strategy, as will be explained later on in the paper, is to generate association rules for each class in the training set separately. This strategy has advantages and disadvantages. The advantage is that with unbalanced training sets (i.e. training sets with rare classes) the small classes do not get overshadowed by the large classes, as is the case with other associative classification approaches. On the other hand, small classes end up generating a huge number of rules since, as will be explained later with the association rules, every feature in the few observations representing the rare classes becomes locally frequent and thus generates rules with high confidence. So dealing with rare classes is what initially motivated this work concerning pruning classification rules. However, in order to generalize the concepts, instead of using the rule generation by class using our *ARC-BC* algorithm [2], we use herein our *ARC-AC* [2] classifier which considers all classes together like other associative classifiers in the literature [6, 7].

In this paper we present an approach to prune the large set of classification rules using the rule performance on the training set. We show with progressive pruning techniques how the number of rules is reduced significantly without jeopardizing the accuracy of the overall classifier. In some cases, the accuracy is actually improved.

In the reminder of the paper we will briefly present the concepts related to association rule mining in Section 2 and will illustrate how these can be integrated to generate an associative classifier. In the same section, we will also introduce related work and highlight their different strategies. The rule pruning approaches will be presented in Section 3 and some experimental results will be illustrated in Section 4. Some conclusions are offered in Section 5.

2 Association Rules and Their Integration in Classifiers

The problem of mining association rules over market basket analysis was introduced in [1]. The problem consists of finding associations between items or itemsets in transactional data. The data is typically retail sales in the form of customer transactions, but can be any data that can be modeled into transactions. For example medical images where each image is modeled by a transaction of visual features from the image, or text data where each document is modeled by a transaction representing a bag of words, or web access data where click-stream visitation is modeled by sets of transactions, all are well suited applications for association rules or frequent itemsets.

Formally, the problem is stated as follows: Let $I = \{i_1, i_2, ...i_m\}$ be a set of literals, called items where m is considered the dimensionality of the problem.

Let $\mathcal{D}$ be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. A unique identifier, *TID*, is given to each transaction. A transaction T is said to contain X, a set of items in I, if $X \subseteq T$. An *association rule* is an implication of the form "$X \Rightarrow Y$", where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \emptyset$. An itemset X is said to be *frequent* if its *support s* is greater or equal than a given minimum support σ. The rule $X \Rightarrow Y$ has a *support s* in the transaction set $\mathcal{D}$ if $s\%$ of the transactions in $\mathcal{D}$ contain $X \cup Y$. In other words, the support of a rule is the probability that X and Y hold together in $\mathcal{D}$. It is said that the rule $X \Rightarrow Y$ holds in the transaction set $\mathcal{D}$ with *confidence c* if $c\%$ of transactions in $\mathcal{D}$ that contain X also contain Y. In other words, the confidence of the rule is the conditional probability that the consequent Y is true under the condition of the antecedent X. The problem of discovering all association rules from a set of transactions $\mathcal{D}$ consists of generating the rules that have a *support* and *confidence* greater than given thresholds. These rules are called *strong rules.*

The first reference to using association rules as classification rules is credited to [4] while the first classifier using these association rules was CBA introduced in [7] and later improved in CMAR [6], and ARC-AC and ARC-BC [9]. The idea is relatively simple. Given a training set modeled with transactions where each transaction contains all features of an object in addition to the class label of the object, we can constrain the association rules to always have as consequent a class label. In other words, the problem consists of finding the subset of strong association rules of the form $X \Rightarrow C$ where C is a class label and X is a conjunction of features (feature set). The difference between CBA, CMAR and ARC-AC and ARC-BC lies in the strategy for rule selection during the classification. They also have some differences in pruning rules. CBA ranks all discovered rules by *precedence* ordering (using confidence then support) and simply selects the first ranked rule that applies given an object to classify [7]. CMAR takes all rules that apply within a confidence range and selects from this set the one with the highest χ^2 measure. ARC-AC and ARC-BC also take all rules that apply within a confidence range, but instead, calculate the average confidence for each set of rules grouped by class label in the consequent and select the class label of the group with the highest confidence average. Another difference is that ARC-AC, CMAR and CBA generate the association rules from all training transactions together. ARC-BC, on the other hand, generates association rules for transactions grouped by class label, each class at a time, giving this way a chance to small classes to have representative classification rules. Another interesting but not very convincing approach proposed in [5] suggests to consider the size of the antecedent and favour long rules before making an allowance for confidence and support. Their experimental results are unfortunately not compelling.

3 Pruning Rules

As stated in [4, 6, 7], associative classifiers generate an overwhelming number of classification rules and it is very important to prune the rules to make the classifier effective and more efficient. We argue that pruning is also very important in order to allow domain experts to tune a classifier by editing rules if necessary.

Our previous experiments show that manual alteration of the rules can lead to significant improvement in the classification [3]. The techniques proposed to prune the rules are based on redundancy and noise elimination and precedence ranking. For example contradictory rules such as $X \Rightarrow C1$ and $X \Rightarrow C2$ are eliminated in the case of single class classification. More specific rules are favoured. For example given two rules $R_1 : X \Rightarrow C$ and $R_2 : Y \Rightarrow C$ if both have the same confidence and $X \subset Y$, only R_1 is kept and R_2 is eliminated. Another accepted method of pruning is *database coverage* introduced in [7] and used in [6]. Database coverage consists of going over all the rules and evaluating them against the training instances. Whenever a rule applies correctly on some instances, the rule is marked and the instances eliminated until all training instances are covered. Finally, the unmarked rules are simply pruned.

Our associative classifier ARC-AC uses only database coverage because other prunings influence the accuracy on many real application datasets. While many rules are eliminated this way, we still find that the number of remaining rules is crushing and further pruning is required. The question is how can we remove more rules without jeopardizing the accuracy of the classifier. We propose to study the performance of each rule in re-classifying the training set and plotting the graph for correct classifications and incorrect classifications for each rule. Figure 1 shows an example. Each rule is plotted with the number of true positives and false positives scored on the training set. The rules plotted high on the graph incorrectly classified many instances. Rules that are plotted towards the right of the graph correctly classified many instances. Note that correct classification does not exclude incorrect classification. One given rule can do both for a large number of instances. The idea of exploiting the graph is to identify culprits of many misclassifications. To do this, we suggest four alternatives that can be executed progressively. Figure 1 illustrates these alternatives. We can visually identify the good and the poor rules.

1. Eliminate the high offender rules: By tracing a horizontal line at a given threshold, we can eliminate all the rules above the line. We suggest a line at 50% by default but a sliding line can also be possible aiming at a certain percentage of rules to eliminate.
2. Eliminate the rules that misclassify more than they classify correctly: By tracing a diagonal line such rules can be identified. Notice that when the axes of the plot are normalized, the diagonal indicates the rules that correctly classify as many times as they misclassify. When the axes are not normalized, the diagonal indicates a relative ratio, which we advocate.
3. Elimination by quadrant slicing: The plot could be divided into four regions. The top left ($RegionA$) contains rules that are incorrect more than they are correct. The top right ($RegionB$) contains rules that are frequently used but equally misclassify and correctly classify. The bottom left ($RegionC$) has rules that are infrequently used but equally misclassify and correctly classify. Finally, the bottom right ($RegionD$) contains the good rules which frequently classify correctly but seldom misclassify. The idea is to successively remove the rules that are in $RegionA$, then $RegionB$, then $RegionC$.

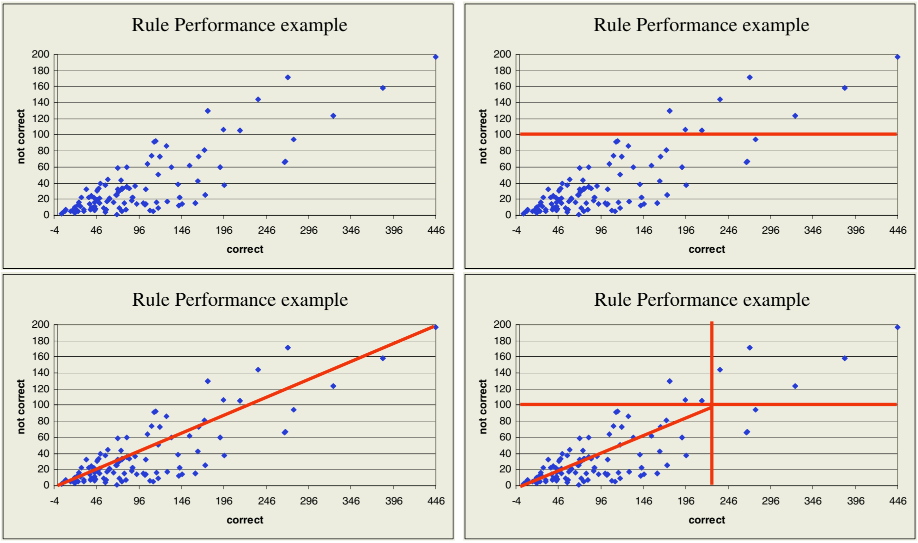


Fig. 1. Filtering by quadrant and diagonal slicing

4. A combination of the above methods: After removing regions A and B, eliminating the rules in $RegionC$ (bottom left) can be costly because many rules may be seldom used but have no replacements. Once removed, other rules are "forced" to play their role and can in consequence misclassify. The idea is to use a diagonal line to identify within $RegionC$ the rules that misclassify more than they are correct. This strategy is a good compromise.

Pruning classification rules is a delicate enterprise because even if a rule misclassifies some objects, it has a role in correctly classifying other objects. When removed, there is no guarantee that the object the rule used to correctly classify will be correctly classified by the remaining rules. This is why we advocate the progressive strategies depending upon the datasets at hand. We found that Strategy 4 (combining quadrant and diagonal pruning) allows in general a good result. A good result here means that we reduce the number of rules while keeping or improving the accuracy of the classifier as much as possible.

4 Experimental Results and Performance Study

We run a battery of experiments to evaluate our strategies. For lack of space, we report herein a representative fraction of these experiments with good and less encouraging results. We used datasets from the UCI ML repository [8] and the performance of CBA and CMAR are from their respective authors' papers [6, 7]. We used a 10 fold cross-validation method for each dataset and what is reported are averages. The table in Figure 3 shows the comparative results for five datasets namely *Breast, Diabetes, Iris, Led7* and *Pima*. We first tested our strategies without any additional pruning then added the best strategy to the

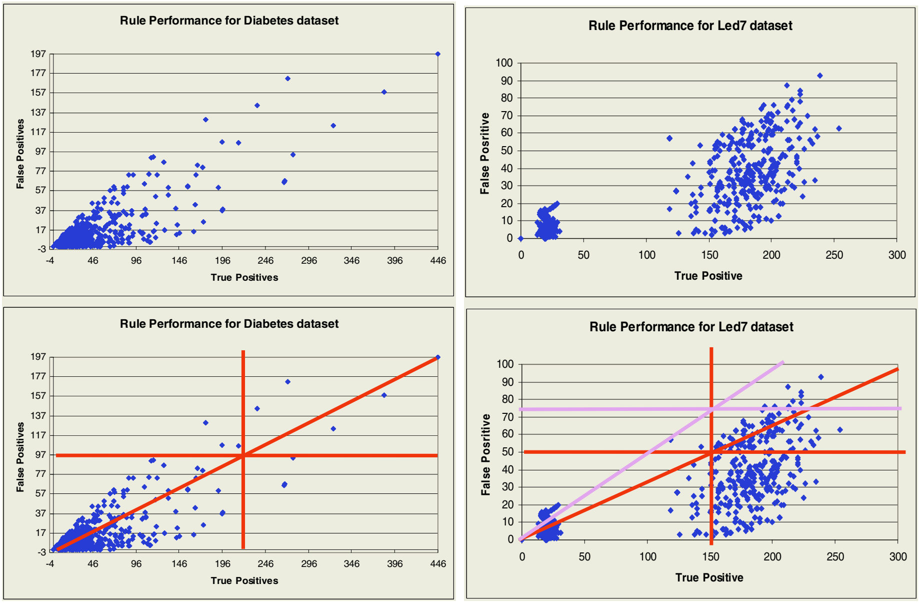
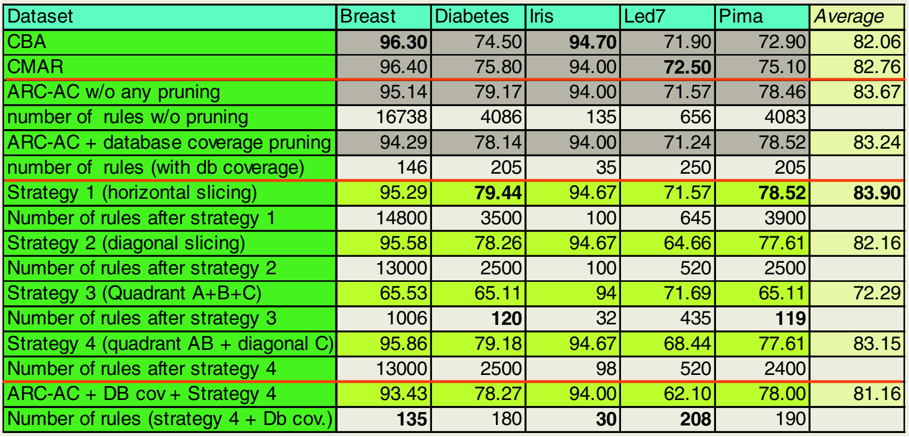


Fig. 2. Rule Performance Plot and Rule pruning for *Diabetes* and *Led7* datasets

database coverage technique. We chose to report on datasets that have different distributions of rules. For instance Figure 2 shows the plots for *Diabetes* with rules concentrating in the bottom left corner and distributed sparsely along the diagonal, and *Led7* with two clusters of rules, one of rarely used rules and one of frequently used rules. The table in Figure 3 first compares ARC-AC with and without database coverage pruning against CBA and CMAR. ARC-AC is the winner on this small collection. With database coverage, the accuracy is still very good while the number of rules drops significantly. Figure 3 also shows the effect of the pruning strategies when no other pruning technique is applied. While Strategy 3 has the worst accuracy result overall, it drastically reduces the number of classification rules without bringing the accuracy too low. In the case of *Led7*, this strategy was actually very good. By eliminating the entire cluster of rules in *RegionC* the performance was better than eliminating a portion of it in Strategy 4. This is because when the entire cluster is eliminated, the remaining rules sharing the task of classifying objects, normally classified by a rule from the cluster, do an excellent job at it. When removing only part of the cluster in Strategy 4, the rules that fire for the objects classified by the pruned rules are actually those remaining from the small cluster and they misclassify indeed. Strategy 3 is thus too drastic, while Strategy 4 is a good compromise.

The winning strategy overall (for the reported datasets) is the simple horizontal slicing of Strategy 1. It outperforms CBA and CMAR on two datasets. In most datasets the bar was put at 50% except for *Led7* for which it was set at 75% since many of its rules are in *RegionB*. However, based on our other evaluations, Strategy 4 is typically the winner overall. By slicing horizontally



Dataset	Breast	Diabetes	Iris	Led7	Pima	*Average*
CBA	**96.30**	74.50	**94.70**	71.90	72.90	82.06
CMAR	96.40	75.80	94.00	**72.50**	75.10	82.76
ARC-AC w/o any pruning	95.14	79.17	94.00	71.57	78.46	83.67
number of rules w/o pruning	16738	4086	135	656	4083	
ARC-AC + database coverage pruning	94.29	78.14	94.00	71.24	78.52	83.24
number of rules (with db coverage)	146	205	35	250	205	
Strategy 1 (horizontal slicing)	95.29	**79.44**	94.67	71.57	**78.52**	**83.90**
Number of rules after strategy 1	14800	3500	100	645	3900	
Strategy 2 (diagonal slicing)	95.58	78.26	94.67	64.66	77.61	82.16
Number of rules after strategy 2	13000	2500	100	520	2500	
Strategy 3 (Quadrant A+B+C)	65.53	65.11	94	71.69	65.11	72.29
Number of rules after strategy 3	1006	**120**	32	435	**119**	
Strategy 4 (quadrant AB + diagonal C)	95.86	79.18	94.67	68.44	77.61	83.15
Number of rules after strategy 4	13000	2500	98	520	2400	
ARC-AC + DB cov + Strategy 4	93.43	78.27	94.00	62.10	78.00	81.16
Number of rules (strategy 4 + Db cov.)	**135**	180	**30**	**208**	190	

Fig. 3. Comparison of CBA, CMAR, ARC-AC and the pruning strategies

at a given percentage of the *False Positives* and then vertically at a certain percentage of the *True Positives*, we generate four regions of unequaled areas A, B, C and D. Then by removing rules in A, B and above the diagonal of C we make sure that we eliminate the rules with the highest ratio of incorrect versus correct classifications, yielding a smaller set of classification rules but a good overall accuracy. Combining Strategy 4 with database coverage further reduces the number of rules while the performance in accuracy remains adequate. In the case of *Diabetes* the accuracy actually improved while the number of rules was reduced by 10%.

5 Conclusion and Future Work

Associative classifiers by piggybacking on the association rule mining technology are cursed by the combinatorial explosion in the number of classification rules generated. This extraordinary number of rules has a consequence on the efficiency of a classifier, but more seriously makes it impossible to manually edit and improve the rules by adding domain knowledge in the model. Inserting domain knowledge is often desirable and association rules are in theory readable by humans. In this paper we propose some strategies to prune the classification rules without severely hindering on the classifier's performance, and sometimes even improve its accuracy. The pruning strategies are simple and are based on individual rule performance when re-classifying the training set. A visual and interactive user application for rule pruning is desired. We are currently working on such interface using the plot presented above that allows interactive selection of rules for pruning and editing, visualizing rule performance and colour coded confidence and support.

References

1. Agrawal, R., Srikant, R. Fast algorithms for mining association rules. In *Proc. Intl. Conf. on Very Large Data Bases*, (1994) 487–499
2. Antonie, M.-L., Zaïane, O. R. Text document categorization by term association. In *Proc. of the IEEE International Conference on Data Mining (ICDM'02)*, (2002) 19–26
3. Antonie, M.-L., Zaïane, O. R., Coman, A. Associative Classifiers for Medical Images, In Mining Multimedia and Complex Data (LNAI 2797), Springer-Verlag, (2003) 68–83
4. Bayardo, R.: Brute-force mining of high-confidence classification rules. In *3rd Intl. Conf. on Knowledge Discovery and Data Mining (KDD'97)*, (1997) 123–126
5. Coenen, F., Leng, P. An evaluation of approaches to classification rule selection. In *IEEE International Conference on Data Mining (ICDM'04)*, (2004) 359–362
6. Li, W., Han, J., Pei J.: CMAR: Accurate and efficient classification based on multiple class-association rules. In *IEEE International Conference on Data Mining (ICDM'01)*, (2001) 369–376.
7. Liu, B., Hsu, H., Ma, Y.: Integrating classification and association rule mining. In *4th Intl. Conf. on Knowledge Discovery and Data Mining (KDD'98)*, (1998) 80–86.
8. UCI repository, http://www.ics.uci.edu/~mlearn/MLRepository.html
9. Zaïane, O. R., and Antonie, M.-L. Classifying text documents by associating terms with text categories. In *Proc. of the Thirteenth Australasian Database Conference (ADC'02)*, (2002) 215–222.

Using Artificial Neural Network Ensembles to Extract Data Content from Noisy Data

Szymon K. Szukalski, Robert J. Cox, and Patricia S. Crowther

School of Information Sciences and Engineering
University of Canberra, ACT 2601, Australia
szymon.szukalski@amused.net, Robert.Cox@canberra.edu.au,
trishc@webone.com.au

Abstract. We have developed a technique to extract points that contain information from a sea of noisy data using an ensemble of Artificial Neural Networks. The technique is relatively simple to use and by using artificial data sets we demonstrate that it can extract a subset of the data that in effect has a higher signal to noise ratio than the original data. We assert that this technique is of practical use in the area of classification, although it does appear to lose points, particularly near the discriminator.

1 Introduction

Artificial Neural Networks (ANNs) are used as classifiers in many problem domains, such as medical [1], industrial [2] and image processing [3], among others. In some cases the data sets used to train the networks are of relatively low quality and may have a considerable number of data points that have little to do with the actual classification problem. These points may be quite numerous and make the learning process for the ANN problematic at best.

As ensembles of networks have been shown to classify data more accurately than a single classifier [4], [5], we examine the performance of ensembles of multi-layer perceptron (MLP) neural networks to see if they can extract useful information from artificially generated data sets with large amounts of noisy data.

2 Previous Work

A previous study [6] looked at voting ensembles, demonstrating that an ensemble of Artificial Neural Networks had better classification power when more of the networks agreed. A subsequent study [7] proposed that data points agreed upon by a significant number of networks in an ensemble could possibly contain a higher information content than those where the networks disagree with each other.

3 Hypothesis

We hypothesise that when a large number of ANNs in a bagging ensemble agree on the classification of a data point, then that data point is more likely to contain useful

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 974–980, 2005.

classification information. A cluster of such points could represent a sub-domain in which the data points contain more data than the data points in the domain as a whole.

4 Experimental Method

We have developed a technique for extracting points with information content from a sea of noisy data. By generating the data ourselves, we know which data has information content and which contains noise.

We generate a 'noise' data set (part A) as shown in figure 1 (left hand side). The data set is two-dimensional within problem domain such that $0 \le x, y \le 1$.

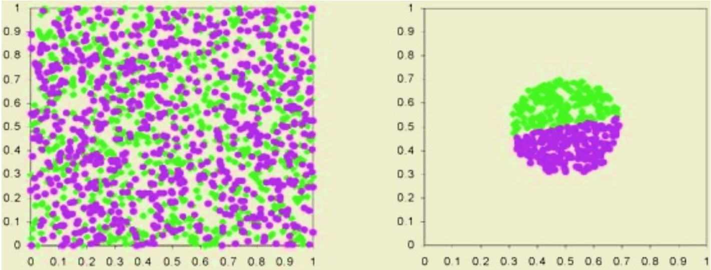


Fig. 1. Left hand side: Part A. Example Noise data set with 1600 points, randomly distributed. Right hand side: Part B. Example 'Good' point sub domain data set with 400 points which are in a specific sub domain and have a linear discriminator

Next we generate a 'patch' of good data (part B) in a sub domain of part A, as shown in the right hand side of figure 1. This patch of good data contains a discriminator. The sub domains and discriminators varied from experiment to experiment. In figure 1 part B, the sub domain is a circle about the point (0.5, 0.5) with radius 0.2; the discriminator follows the linear equation:

$$y = \frac{2}{5} + \frac{x}{5} \tag{1}$$

Two other discriminators are used in subsequent experiments

$$y = \frac{1}{2}\sin\frac{3\pi x}{2} + \frac{1}{2} \tag{2}$$

$$y = \frac{1}{2}\sin 6\pi x + \frac{1}{2} \tag{3}$$

The two data sets are merged to form a combined data set of 2000 points (figure 2), giving a signal to noise ratio for the total problem domain in this data set of 1:4. This method produces the starting point data set for each experiment.

The combined data set is then randomly split into four subsets each of 500 points. These sub sets are called Train, Test, Validation1 and Validation2. Validation2 is used to measure the accuracy of the ensemble; it was within an acceptable range for all experiments, details are not included due to space limitations.

An ensemble of 100 networks is created using bagging [8], with the training and testing data sets being re-sampled for each network within the ensemble. An automated training algorithm [9], [10] ensures that the networks are trained to an appropriate level of generalization, based on the Test subset.

The results of each network within the ensemble are examined using the Validation1 data set, to calculate the number of networks that agree on each data point.

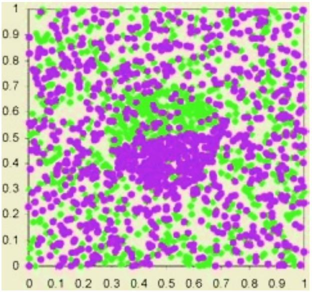


Fig. 2. Combined Dataset

We then plot the percentage of networks that agree on points versus the accuracy of the points that they agree on (as measured by the Validation 1 data set). Figure 3 has the output from our base line experiment. When we graph this data we remember the peak point - the number of networks that produced the highest accuracy in the points that they agree on. In the experiment graphed in figure 3 the peak is reached at 92 networks.

We also plot the number of points *in* the sub domain of the patch of good data that the networks agree on, versus the percentage of networks that agree on the points (figure 4). On the same axis we plot the number of points *outside* the good data patch sub domain that the networks agree on. Where the lines cross over is the point that the signal to noise ratio is 1:1 (i.e. as many points are in the sub domain as are outside it). Because our data set is two-dimensional we can plot the actual points that the ensemble agrees upon at the peak point (the top of the graph in figure 3 – 92% in this case). An example of this is shown in figure 5. As can be seen, the points that remain are predominantly inside the patch of good data. This supports our hypothesis that when the ensemble agrees on the classification of a point, then that point has more information content than the data set as a whole.

In all we conducted 11 different experiments, one of which was a control where the ‘patch’ contained random data.

5 Results from the 11 Experiments

The results from the 11 experiments are shown in tables 1 and 2. The most significant columns are the percentage in sub-domain at start and percentage in sub-domain at peak.

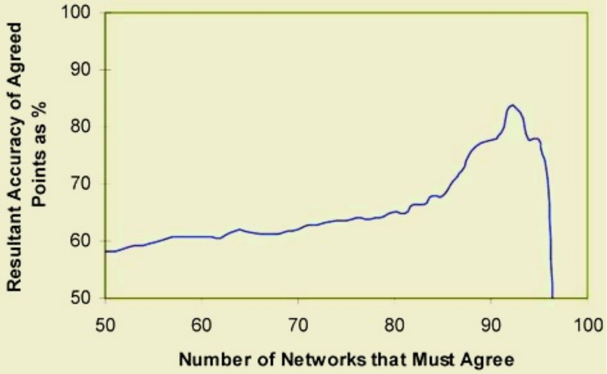


Fig. 3. Number of networks that agree vs the accuracy of the points they agree on

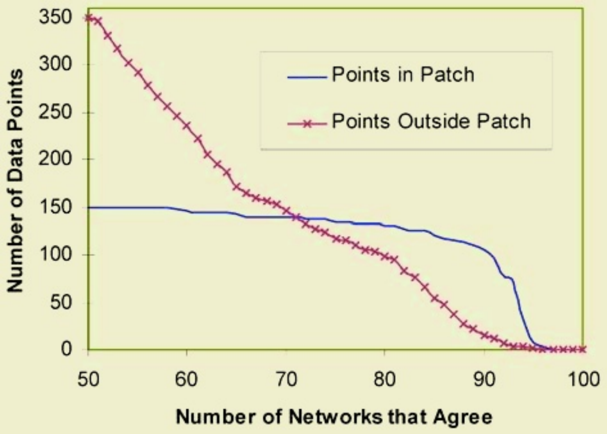


Fig. 4. Number of points inside the 'good' data sub domain (patch) and number of points outside the 'good' data sub domain (patch) plotted against number of network that agree on the points

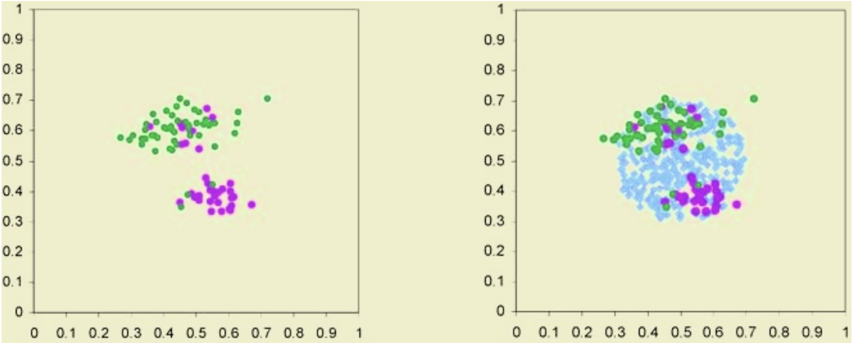


Fig. 5. The data points agreed on by 92 networks. The Right hand side shows the data points against the original patch of 'good' data

Table 1. Results from 11 experiments showing the accuracy at the peak, the number of networks that agree and the number good points at the peak based on Validation1 (500 points)

ID	Accuracy at Peak	# Networks Agree at Peak	# Points Agree at Peak	Initial data inside patch	Initial Data outside Patch	Points in Patch at Peak	Points outside Patch at Peak
1	73.08%	100	182	152	348	102	80
2	85.53%	92	85	150	350	78	7
3	90.00%	100	30	147	353	30	0
4	71.74%	95	138	197	303	127	11
5	67.06%	87	252	262	238	247	5
6	75.00%	99	4	185	315	4	0
7	68.42%	100	190	152	348	126	64
8	61.27%	98	346	154	346	129	217
9	69.14%	100	175	213	287	153	22
10	55.91%	67	499	213	287	213	286
11	57.14%	91	21	150	350	16	5

Table 2. Distribution of data in patch, discriminator, and percentage of data in sub-domain before by experiment

ID	Patch Distribution	Discriminator formula	% in sub-domain at start	% in sub-domain at peak	Notes
1	C2	(1)	30.4%	56.0%	Note 1
2	C2	(1)	30.0%	91.8%	
3	C0	(1)	29.4%	100.0%	
4	C8	(1)	39.4%	92.0%	
5	Q2	(1)	52.4%	98.0%	
6	Q3	(1)	37.0%	100.0%	
7	C2	(2)	30.4%	66.3%	
8	C0	(2)	30.8%	37.3%	
9	C8	(2)	42.6%	87.4%	
10	C8	(3)	42.6%	42.7%	Note 3
11	C2	None	30.0%	76.2%	Note 2

Note 1 – This experiment did not use bagging.
Note 2 – this is a ‘control’ experiment.
Note 3 – This data set has a quite complicated discriminator function.

The codes for the patch distributions are as follows:

C0 - Circle radius 0.2 about the point (0.6, 0.6).
C2 - Circle radius 0.2 about the point (0.5, 0.5).
C8 – Circle radius 0.3 about the point (0.6, 0.6).
Q2 – Horizontal rectangle where x is in the range 0 – 1 and y < 0.4.
Q3 – Horizontal rectangle where x is in the range 0 – 1 and y > 0.4 and y < 0.6.

Experiment 1 was an ensemble that did not use the bagging technique, rather we randomized the initial weights in the network.

Experiments 2 through 10 all used bagging ensembles with varying distributions and discriminators. In all cases within experiments 2-10, some improvement is made in the ratio of points that include data.

Experiment 6 does not appear to have performed particularly well, finding agreement on only four points at its highest accuracy, where 99 networks agree. However, for 98 networks agreeing in this experiment, 27 data points were found, all of which were inside the ‘good’ data patch. In experiment 6, the good data was in a long thin rectangle with the discriminator a diagonal across it; a quite difficult problem, so it is not surprising that the ANNs didn’t easily agree on points.

In experiment 10 the discriminator function was quite complicated and classification appeared to be beyond the capabilities of the neural network training algorithm, even here, the one point that was removed was outside the patch.

Note also that experiment 11 is a control experiment where no meaningful data was included in the patch, although the data density is still higher that of the surrounding area. Experiment 11 shows that when the number of points agreeing and the accuracy at the peak value are both low, there may be no meaningful data content at all. The graph of resultant accuracy of agreed points was uncharacteristically flat with a sharp peak, whereas for all the other experiments the rise to the peak was quite gradual, in line with that shown by figure 3.

6 Conclusions and Future Directions

An ensemble of bagged networks configured as a filter using the method we have described is able to improve data quality. The notable disadvantage of this technique is that a fair amount of data is lost, including data points around the discriminator. The technique appears to be suitable for data sets that have a low inherent accuracy but a large amount of data. In such a data set, the technique will produce a subset of the data that has higher information content than the original data set. The value of this smaller data set is the subject of further research. We have already used the method successfully as a front end in a technique called RDC-ANNE [11] in which a neural network ensemble preprocesses data that is subsequently fed to a decision tree. Future work will involve further analysis of the usefulness of these subsets of data.

References

1. Shadabi, F., Cox, R., Sharma, D. & Petrovsky N. Experiments with a Neural Network Ensemble to Predict Renal Transplant Outcomes. Proc AISAT 2004 The 2nd International Conference on Artificial Intelligence in Science and Technology, Hobart, Australia, 21-25 November 2004 pp 271-276
2. Fogelman-Soulie, F. & Gallinari, P. (ed) Industrial Applications of Neural Networks. World Scientific Publishing Co. (1998)
3. Egmont-Petersen, M., de Ridder, D. & Handels, H. Image Processing with Neural Networks – A Review. Pattern Recognition 35 (2002) pp2279-2301
4. Brieman, L.: Bagging Predictors. Machine Learning 24(2) (1996) 123-140.
5. Clemen, R. Combining Forecasts: A Review and Annotated Bibliography. Journal of Forecasting 5 (1989) 559-583.
6. Cox, R., Clark, D. & Richardson, A. An Investigation into the Effect of Ensemble Size and Voting Threshold on the Accuracy of Neural Network Ensembles. Lecture Notes in Computer Science vol. 1747. Proceedings of the 12th Australian Joint Conference on Artificial Intelligence: Advanced Topics in Artificial Intelligence pp268-277 (1999).

7. Clark, David. Using Consensus Ensembles to Identify Suspect Data. Knowledge-Based Intelligent Information Systems 8th International Conference, KES 2004 Wellington, New Zealand, September 2004 Proceedings, Part II pp 483-489, Springer-Verlag 2004.
8. Sharkey, A.J.C., ed.: Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems. Perspectives in Neural Computing. Springer-Verlag, London (1999)
9. Crowther, Patricia, Cox, Robert & Sharma, Dharmendra. A Study of the Radial Basis Function Neural Network Classifiers using Known Data of Varying Accuracy and Complexity, Knowledge-Based Intelligent Information Systems 8th International Conference, KES 2004 Wellington, New Zealand, September 2004 Proceedings, Part III p 210-216, Springer-Verlag 2004.
10. Cox, R. J. & Crowther, P. S. An Empirical Investigation into the Error Characteristics of Neural Networks, Proceedings AISAT 2004 The 2nd International Conference on Artificial Intelligence in Science and Technology, Hobart, Australia, 21-25 November 2004 p 92-97.
11. Shadabi, F., Cox, R., Sharma, D. & Petrovsky, N. A Hybird Decision Tree – Artificial Neural Networks Ensemble Approach for Kidney Transplantation Outcomes Prediction. Submitted to KES 2005.

Identification of a Motor with Multiple Nonlinearities by Improved Genetic Algorithm

Jung-Shik Kong[1] and Jin-Geol Kim[2]

[1] Dept. of Automation Eng., Inha University, YongHyun-Dong, Nam-Gu, Inchon, Korea
selkirk@nate.com
[2] School of Electrical Eng., Inha University, YongHyun-Dong, Nam-Gu, Inchon, Korea
john@inha.ac.kr

Abstract. This paper presents a mathematical model that employs a new genetic algorithm for motor identification. Mechanical structures require precise motor information for high control performance. However, it is difficult to acquire accurate motor information and a genetic algorithm can be an adequate method to search unknown parameters using only angular position. The previous methods by using conventional genetic algorithms do not give the most optimal result since they cannot adjust the parameters with infinite precision. A new method is needed to identify uncertain motor information. This paper proposes a mathematical model that was searched by the newly proposed genetic algorithm. The induced motor model is verified through the real experiment.

1 Introduction

In control engineering, a precise mathematical model of a motor assists the designers in making high performance controllers. Also, if a motor model is accurate, the optimal solution from a simulation can be immediately applied to the real world. However, an incorrect motor model cannot apply directly in the real system. It means the error between simulation and experiment. When a motor is used alone in the absence of a mechanical structure, its specification from the manufacturers is accurate. But, the majority of motor systems are employed in various structures with uncertainty. Thus, it is hard to apply the basic motor system because of nonlinearity and dynamic characteristics of external structures. Users that employ this module are unaware of exactly how the motor model behaves through the entire system. The nonlinearity of the system causes a limit cycle and an instability leading to inaccurate information regarding the motor and structure. So, simulation results are incapable of describing the real system as a result. A real system model is prerequisite needed in applying an accurate simulation result to the real world. The most powerful method for searching a mathematical model is generally known as system identification. System characteristics are searched by system identification using from observing the input and output of the system. System identification employs various methods; such as neural network [1-2], fuzzy algorithm [3-5], and genetic algorithm [6-9] etc. Each algorithm has its own benefits. Here, the genetic algorithm is used to search for an optimal solution through boundary. The genetic algorithm, due to its benefits, is used in a variety of different fields [10].

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 981–987, 2005.

In spite of these merits, previous genetic algorithms have some limitations. Usually if the gene is used as binary code, it will not be able to search within a gap. In the case of the popular decimal code, the genetic algorithms' searching performance deteriorates with increasing generation numbers. General decimal-coded genetic algorithms always perform searching for optimal data within a whole [11]. In case of previous decimal-coded genetic algorithm, fitness value is settled down when some number of generations is passed. So, if possible, it is required to reduce boundaries of crossover and mutation in order to improve searching ability.

In this paper, a motor identification is performed using an improved genetic algorithm. Its application requires a sigmoid function to be applied at the crossover sequence. This sigmoid function makes the genetic algorithm search in greater detail. Thus, the improved genetic algorithm enhances searching capability. Finally it is shown that the motor model is identified with the improved genetic algorithm using the data of input and output from the given system. Simulation and experiment results are shown in various situations.

2 Motor System Modeling

Block diagram of a motor system is represented in Fig. 1. Most of the motor systems include the linear and the nonlinear part. The motor system has two parts. One is linear part and the other is nonlinear part that includes voltage saturation, backlash from reducer and so on. Here, nonlinearity might cause a limit cycle or influence system dynamics.

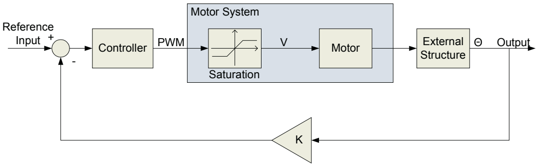


Fig. 1. Block diagram of a motor system

A motor transfer function, between voltage of the motor and the angle of joint is shown in Fig. 1. This is defined as:

$$\frac{\theta(s)}{V(s)} = \frac{K_a}{L_a J_m s^3 + (L_a f_m + R_a J_m)s^2 + (R_a f_m + K_a K_b)s} \tag{1}$$

where $V(s)$, $\theta(s)$, L_a, R_a, J_m, f_m, K_a, and K_b are the input voltage, output angle, armature inductance, armature resistance, moment of inertia of the motor, viscous-friction coefficient of the motor, the motor-torque proportional constant, and a proportional constant from back electromotive force, respectively.

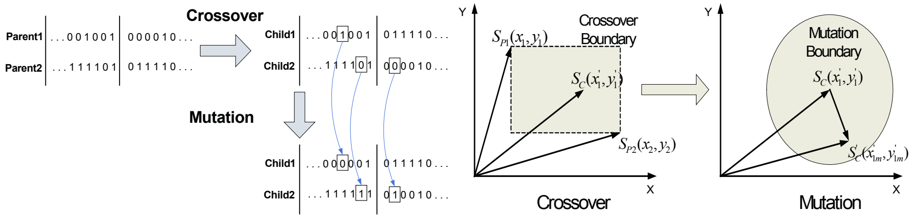
The saturation shown in Fig. 1 represents the maximum voltage from the hardware. This is caused by the maximum driving capability of motor power driver. The saturation parameters can be expressed as:

$$V = \begin{cases} V_{Max} & P \geq V_{Max} \\ V_{Min} & P \leq V_{Min} \\ P & otherwise \end{cases} \tag{2}$$

3 Identification Problems

3.1 Genetic Algorithm

A genetic algorithm is a kind of searching algorithm that it can find the optimal solution without solving a differential equation. Previous studies using the genetic algorithm are separated into two sections. One is a binary-coded genetic algorithm and the other is a decimal-coded genetic algorithm. The binary-coded genetic algorithm uses binary code during crossover and mutation processes. It has the advantage of easily being able to apply genetic concepts. However, the own accuracy of algorithms limits the searching precision for the optimal solution. The other method using a decimal-coded genetic algorithm applies real values when this algorithm executes crossover and mutation. It can search a more precise optimal solution than the binary-coded genetic algorithm because of using real values. However, this algorithm is dictated by time, with the number of generations settling down with an increase in time. Although the fitness value is settled down, it is hard to say that the result from decimal-coded genetic algorithm reaches the optimal point. Fig. 2 represent crossover and mutation processes of binary-coded genetic algorithm and decimal-coded genetic algorithm, respectively.



(a) Binary-coded GA (b) Decimal-coded GA

Fig. 2. Crossover and mutation of binary-coded GA and decimal coded GA

In Fig. 2(b), $S_{P1}(x_1, y_1)$ and $S_{P2}(x_2, y_2)$ are set of parents, $S_C(x_1^{'}, y_1^{'})$ is the result of crossover and $S_C^{'}(x_1^{'}, y_1^{'})$ is the output of mutation. Here, x and y represent by the parameters to find out by using decimal-coded genetic algorithm.

To improve the problems of previous algorithm, more advanced algorithm is proposed. It uses a decimal-coded genetic algorithm basically. As shown in Fig. 2(b), decimal-coded genetic algorithm has the maximum crossover boundary between two parents during the crossover. In conventional decimal-coded genetic algorithm, the fitness function is settled after passing the generation count and the optimal data is searched at this time. The data, which stay close to the optimal value, may not technically be the desired data. If crossover and mutation boundary is reduced while generation count has passed, more optimal solution can be searched. To attack this problem, a sigmoid function is used in crossover and mutation. Equations (3-4) show the suggested crossover and mutation.

$$
\begin{aligned}
x_1^{'} &= x_1 + \delta(t)\times(1-\lambda_1)x_2, \quad x_2^{'} = x_2 + \delta(t)\times(1-\lambda_2)x_1 \\
y_1^{'} &= y_1 + \delta(t)\times(1-\lambda_3)y_2, \quad x_2^{'} = y_2 + \delta(t)\times(1-\lambda_4)y_1
\end{aligned}
\tag{3}
$$

$$
\begin{aligned}
x_{1M}^{'} &= x_1^{'} + \delta(t)\times \mathrm{sgn}(\alpha)\times d\times\lambda_m \\
y_{1M}^{'} &= y_1^{'} + \delta(t)\times \mathrm{sgn}(\alpha)\times d\times\lambda_m
\end{aligned}
\qquad \delta(t) = 1-\frac{1}{1+e^{-(10t+5)}}
\tag{4}
$$

where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are the random number ($\lambda_{1,2,3,4} \in (0,1)$) and t is generation count when genetic algorithm is conducted. By using $\delta(t)$, crossover and mutation boundary is reduced according to the time. Fig. 3 depicts comparison average and maximum of fitness values.

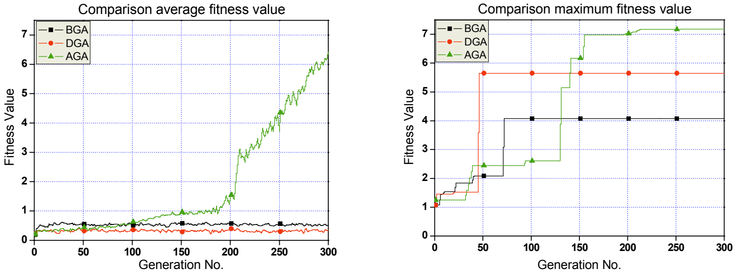


Fig. 3. Comparison of average and maximum fitness

Here, BGA means binary-coded genetic algorithm, DGA stands for decimal-coded genetic algorithm, and AGA is the acronym of advanced decimal-coded genetic algorithm. From Fig. 3, average fitness from proposed genetic algorithm has a relatively high fitness value. As it is well known from identification results, the larger a fitness value, the larger the similarity to real motor models.

3.2 Identification of a Motor System

The proposed identification method is shown in Fig. 4. Plant system block $G(s)$ represents a real plant, while $\hat{G}(s)$ denotes a mathematical model controlled by genetic algorithm.

The mathematical model is altered by a genetic mechanism in the following 3 steps. First, genetic mechanism compares with results between the plant system and mathematical model. It then compares fitness values between new generated maximum data and previous maximum data. Finally, the mathematical model is changed if the new fitness value is larger than the previous fitness value. These steps are continued until the maximum generation is reached.

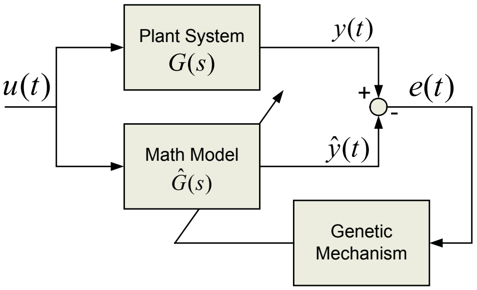


Fig. 4. Block diagram of system identification

4 Experiment

The experimental system only consists of a motor and a reducer. Table 1 shows the motor specifications. Its mathematical model is described as:

$$T_M = \frac{110464760.272}{s^3 + 16340.322s^2 + 1992784.27530s} \tag{5}$$

Table 1. Motor specification

Symbol	Value	Unit	Symbol	Value	Unit
L_a	3.629×10^{-4}	H	K_b	0.01804	Vs/rad
R_a	5.93	Ω	f_m	4.5×10^{-7}	Kgm^2
K_a	0.01804	Nm/A	J_m	1.0472×10^{-7}	Nm

Friction and moment of inertia at the reducer are also included in the motor specifications. In addition, applied genetic parameters for the identification of a motor system are shown in Table 2. The P_R of fitness function is the position values of a real system and P_M is the position values of a mathematical model, respectively. The equation (6) shows an identification model of motor system by genetic algorithm.

$$T_I = \frac{9991390}{s^3 + 2718.94s^2 + 223101s} \tag{6}$$

Table 2. Parameter of genetic algorithm

Parameters	Value	Parameters	Value
Generation No.	300	Crossover Rate.	0.6
Population No.	100	Mutation Rate	0.1
Fitness function	$fit = 1/\sum\lvert P_R - P_M\rvert$		

From the model, some differences between simulation and experimental results existed in the motor system. Fig. 5 depicts the step response comparison between math model and identification result. From Fig. 5, the system non-linearity and uncertainty are represented as the differences between the mathematical model by specification and the real motor model. As shown in Fig. 5, our new mathematical model that applied by improved genetic algorithm is close to the real model.

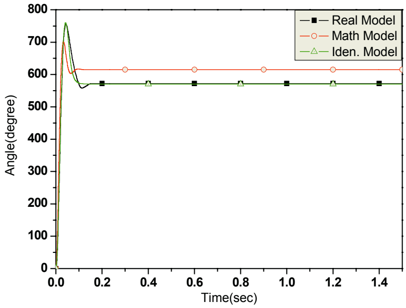
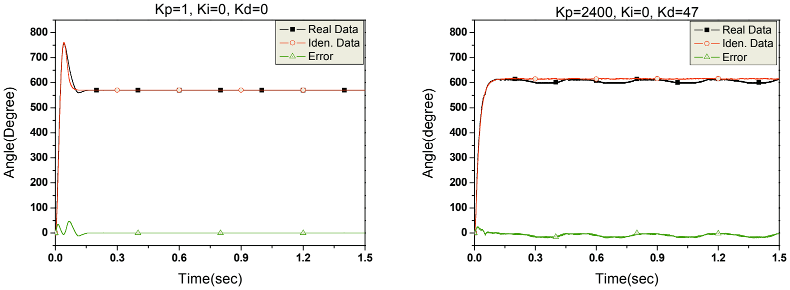


Fig. 5. Comparison of step responses according to model

The identified model is examined with PID controllers in order to verify the validity. Fig. 6 shows a comparison of identification according to the change of the various control gains when the corresponding controllers are applied. As shown in figures, the maximum relative difference between real model and identification results is 7.6% for the case of Fig. 6(a), and relative difference is 4.2% for the case of Fig. 6(b). In these both conditions, the identified model by the proposed genetic algorithm was still found to be similar to the real motor model.



(a) (b)

Fig. 6. Identification results according to PID gains

5 Conclusion

In this paper, motor system identification by using an improved genetic algorithm is suggested. It is hard to use a motor model by only its specification for the control because of its nonlinearity and uncertainty. In order to match the simulations with real world, a more accurate motor model was needed. It is shown in this paper that the proposed genetic algorithm outperforms the previous genetic algorithm and the better mathematical model can be obtained through the simulations.

In the future, a new control algorithm must be used in an attempt to reduce the effects of non-linearity with identification methods. Since it is hard to eliminate the limit cycle with only a PID controller, a more robust control algorithm will be developed by using the approach proposed in this paper, which is applicable to the MIMO system like a humanoid robot with multiple motors.

Acknowledgement

This work was supported by INHA UNIVERSITY Research Grant.

References

1. Lu, S. and Basar, T.: Robust Nonlinear System Identification Using Neural-network Models, IEEE Transactions on Neural networks, Vol. 9 (1998) 407 - 429
2. Mills, P.M. and Zomaya, A.Y.: A Neural Network Approach to On-line Identification of Nonlinear Systems, IEEE int. Conf. Neural Networks (1991) 202-207
3. Vachkov, G. and Fukuda, T.: Identification and Control of Dynamical Systems Based on Cause-effect Fuzzy Models, Int. Conf. IFSA Vol. 4 (2001) 2072-2077
4. Park, M.K., Ji, S.H., Kim, M.J. and Park, M.: A New Identification Method for a Fuzzy Model, Int. Conf. Fuzzy System Vol. 4 (1995) 2159-2164
5. Golob, M. and Tovornik, B.: Identification of Nonlinear Dynamic Systems with Decomposed Fuzzy Models, Int. Conf. System Man and Cybernetics Vol. 5(2000) 3520-3525
6. Akramizadeh, A., Farjami, A.A. and Khaloozadeh, H.: Nonlinear Hammerstein Model Identification Using Genetic Algorithm, Int. Conf. ICAIS (2002) 351-356
7. Mitsukura, Y., Fukumi, M., Akamatsu, N. and Yarnamoto, T.: A System Identification Method Using a Hybrid-type Genetic Algorithm, SICE Vol. 3 (2002) 1598-1602
8. Tipsuwanporn, V., Piyarat, W., Tarasantisuk, C., Suesut, T. and Charean, A.: Identification of Motion Control System Using Genetic Algorithms, Int. Conf. Power Electronics and Drive Systems, Vol. 2 (2001) 854-857
9. Liu Changliang, Liu Jizhen, Niu Yuguang and Yao Wanye : The Application of Genetic Algorithm in Model Identification, TENCON '02 Vol. 3, (2002) 1261-1264
10. Kong, J.S., Lee, B.H. and Kim, J.G.: A Study on the Gait Generation of a Humanoid Robot Using Genetic Algorithm, SICE (2004) 187-191
11. Lee, K.Y. and Mohamed, P.S.: A Real-coded Genetic Algorithm Involving a Hybrid Crossover Method for Power Plant Control System Design, CEC '02 Vol. 2 (2002) 1069-1074

Program Simplification in Genetic Programming for Object Classification

Mengjie Zhang, Yun Zhang, and Will Smart

School of Mathematics, Statistics and Computer Science
Victoria University of Wellington
P. O. Box 600, Wellington, New Zealand
mengjie@mcs.vuw.ac.nz

Abstract. This paper describes a program simplification approach in genetic programming (GP) to the use of simple algebraic techniques, prime numbers and hashing techniques for object classification problems. Rather than manually simplifying genetic programs after evolution for interpretation purpose only, this approach automatically simplifies genetic programs during the evolutionary process. This approach is examined on four object classification problems of increasing difficulty. The results suggest that the new simplification approach is more efficient and more effective than the basic GP approach without simplification.

1 Introduction

Classification tasks arise in a wide variety of practical situations. Diagnosing medical conditions from medical imaging, recognising words in streams of speech, and identifying fraudulent financial transactions are just three examples. Computer based solutions to many of these tasks would be of immense social and economic value. However,writing such computer programs is difficult, time consuming, and often infeasible: human programmers are often unable to identify all the subtle and interrelated conditions that are needed to distinguish between the different classes.

As a promising automatic programming approach, genetic programming (GP) [1] has been applied to object classification and detection tasks since the mid-1990s [2–5]. In many cases, the GP system has achieved reasonable level of success. While showing promise [6–9], current GP techniques are frequently do not give satisfactory results for difficult tasks. One problem is the redundancy of programs. Typically, the programs are not simplified until the end of the evolutionary process, and then only to enable analysis. However, the redundancies also affect the search process. They force the search into exploring unnecessarily complex parts of the search space, which will increase the training time [10, 11]. The redundancies and complexities have the undesirable consequences that the search process is very inefficient, and the programs are very difficult to understand and interpret. However, the redundant components of the evolving programs may play an important role in the evolutionary process by providing a wider variety of possible program fragments for constructing new programs.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 988–996, 2005.

The goal of this paper is to invent a method in GP that does program simplification during the evolutionary process. We will investigate the effect of performing simplification of the programs during the evolutionary process, to discover whether the reduction in complexity outweighs the possible benefits of redundancy. This approach will be examined on four object classification problems of increasing difficulty and compared with the GP method without simplification.

2 Program Simplification

The function set consists of the four arithmetic operators and a conditional operator, which can be grouped into three operator family/categories: + and −, × and %, and *if*. The simplification algorithm developed in this paper addresses simplification of all the categories of functions.

2.1 Simplification of Programs with the Same Operator Family

In the four arithmetic operators, × and + are commutative. Redundancy occurred in evolved genetic programs between × and % or between + and − can be simplified by finding the greatest common sub-multiset (GCS). For example, the fraction representation of a genetic program $\frac{w \times x \times z \times z \times y}{x \times y \times z \times x}$ can be simplified to $\frac{w \times z}{x}$ by finding and removing the GCS $x \times y \times z$:

$$\frac{w \times x \times z \times z \times y}{x \times y \times z \times x} \Longrightarrow \frac{w \times z}{x} \tag{1}$$

Similarly, $(w + x + z + z + y) - (x + y + z + x)$ can be simplified to $(w + z) - x$ by eliminating the GCS $(x + z + y)$.

To reduce the time complexity, we developed a new method, which uses primes and prime products for finding the GCS in two multisets in a program. Using the program illustrated in equation 1 as the example, the algorithm (called *Simp1*) is briefly described as follows.

- Associate each variable (feature terminal) with a unique prime number. In the previous example, we can do: $w \rightarrow w.3$, $x \rightarrow x.5$, $y \rightarrow y.7$ and $z \rightarrow z.11$.
- Group all variables and their primes to form the two multisets. For example, $G_1 = w.3, x.5, z.11, z.11, y.7$ and $G_2 = x.5, y.7, z.11, x.5$.
- Calculate the prime product (PP) for each multiset. $PP_1 = 3 \times 5 \times 11 \times 11 \times 7 = 12705$ and $PP_2 = 5 \times \times 7 \times 11 \times 5 = 1925$.
- Find the GCS. Divide the prime product of the longer set (PP1) by every element prime of the shorter set. If it is dividable, put the element with this prime to GCS and replace the dividend with the quotient; otherwise, skip the element. This process is continued until all elements of the shorter set are tried. In this example, the GCS = {x, y, z}.
- Eliminate the GCS from both sets and simplify the program. In this example, we easily get the simplified program $\frac{w \times z}{x}$ as shown in equation 1.

2.2 Tidy-up Rewriting of Genetic Programs

For most programs/subprograms with the same category of operators, they are not as clean and tidy as the one described in section 2.1 but usually quite messy. For example, the evolved program `(% (× (% (% w x) (% y x)) (% (× z z ) (% z y))) x)` is basically the same as the one presented in equation 1, but it is much messier. To convert this form (original form) of the program into a clean one shown in equation 1, we developed a tidy-up rewriting algorithm.

The basic idea of tidy-up rewriting a genetic program is to use a two-layer storage table for each node in the program tree, as shown in figure 1 (left). We associate each non-terminal node with a two-layer table, one positive layer and a negative layer. The tables are constructed in the bottom up direction for the non-terminal nodes. If the operator of a node is positive such as × and +, the positive layer of its associated table will concatenate all elements stored in positive layer of the child nodes, and the negative layer will concatenate all elements stored in the negative layer of the child nodes. If the operator of the node is negative such as % and −, then the positive layer of its associated table will concatenate all the elements stored in the positive layer in the left child node and those in the negative layer in the right child node. The negative layer will concatenate the elements in the negative layer of the left child node and those in the positive layer of the right child node.

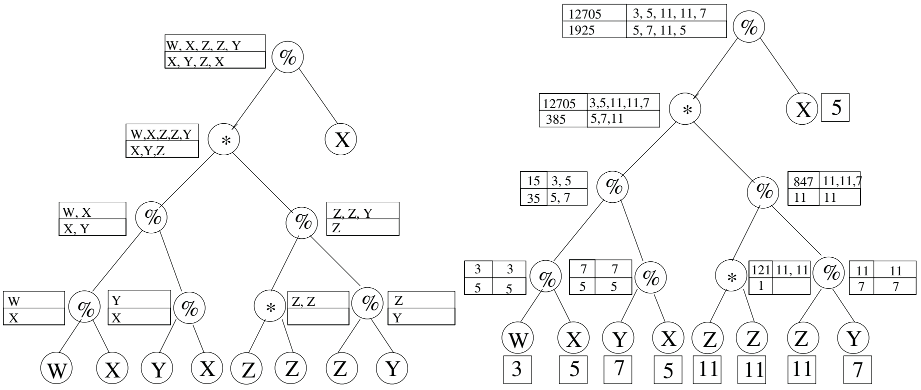


Fig. 1. Tidy-up rewriting of genetic programs

In this way, a double-layer table can be constructed for each of the non-terminal nodes in the program tree. By the end of the process, the table associated with the root node, which is the output of the program, could be easily written in the tidy-up form like the one in equation 1, and accordingly the algorithm *Simp1* can be applied for simplification.

To speed up the process, we replace the elements in the associated tables for the non-terminal nodes in a program tree with their corresponding prime numbers and prime products so that the good properties of primes in algorithm *Simp1* can be adopted, as shown in figure 1 (right).

Double Decker Bus. To further improve the time complexity of the tidy-up process, we developed a new structure, *double-decker bus (DDB)*, to extend the double layered prime tables. This structure considers not only the operators and feature terminals (primes), but also the numeric/constant terminals. The DDB structure has two decks and two wheels, as shown in figure 2. The top deck, called *PosDeck*, contains a set of primes and prime products, corresponding to the positive layers in the double layer tables shown in figure 1. The bottom deck, called *NegDeck*, also contains a set of primes and the product of these primes, but corresponds to the negative layers of the double layer tables shown in figure 1. The front wheel stores the operator family in the function set. Please note that at this stage we assume that the entire program tree must have the same operator family, either all with + and −, or with × and %. Thus, the front wheel stores positive operators such as + or ×, and negative operators such as − or %. The back wheel stores the numeric terminal values, which are floating point constants. These constants are not associated with any primes, and their original values are used in the process.

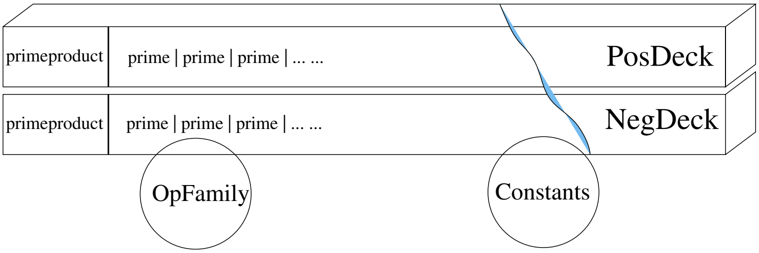

primeproduct
prime | prime | prime | ... ...
PosDeck
primeproduct
prime | prime | prime | ... ...
NegDeck
OpFamily
Constants

Fig. 2. Double decker bus (DDB) – A new structure for tidy-up rewriting of programs

2.3 Simplification of Programs with Different Operator Family

In the actual situations, most genetic programs evolved in the evolutionary process will mix up all the four arithmetic operators even with the conditional operator `if`. Thus, we discuss the case of mixing up the four arithmetic functions in this subsection and the conditional operator in the next.

For example, given the genetic program below:

```
(% (× (× (+ x y) (− z 3) (% y (+ y x))) (× (% (+ z x) w) y) ) )
```

we want to do the tidy-up rewriting to the format of $\frac{(x+y)\times(z-3)\times y\times w}{(z+x)\times y\times(y+x)}$, then perform simplification to the final form of $\frac{(z-3)\times w}{z+x}$. In this program, the main operator family is × and %, but the program also has some *compound terms* such as `(x+y)` and `(z+x)`. For these compound terms, the algorithms discussed so far cannot easily associate a prime number and perform simplification.

Double Hashing. To perform tidy-up rewriting of the program with the above compound terms, we can use an ordered list of prime numbers and assign prime numbers to the compound terms one after another. We can firstly check whether the same compound term has been associated with a prime or not before we associate a prime with a compound term (a subprogram). If it has been already assigned a prime, then use the same prime for that compound term. Only if a compound term has not been assigned any primes, we pick up a new prime number for it. However, this approach is time consuming, particularly when the program is large.

To reduced time, we introduced a double hashing approach to finding appropriate prime numbers for a compound term/subtree in a program. In this approach, a quartuple of elements were collected from the DDB for the sub program trees: the prime product of the PosDeck (*PosPrimeProduct*), the front wheel for the operators (*OpFamily*), the prime product of the NegDeck (*NegPrimeProduct*), and the back wheel for the constant terminals (*Constants*). Then we use two levels of hashing to get a prime number for the subtree. In the first level, the *PosPrimeProduct* and *OpFamily* were used as the keys of the sub program to obtain a slot (an address), from which the second level hashing hashes the *NegPrimeProduct* and *Constants* to produce a prime number. Note that buckets were used in both levels to avoid/reduce potential collisions.

It is important to note that the algorithm *Simp1* is applied to the non-terminal nodes from time to time during the simplification process, in particular when there are mixed up family of operators.

2.4 Simplification of the `if` Function

The `if` function has a different structure from the arithmetic operators. To perform simplification on the `if` function, we treat it as a singleton function, where all children are stored in the *PosDeck* of the DDB and the *NegDeck* would always keep empty. This function is considered a "compound" term and the double hashing approach is always applied to this function to get a prime number associated with.

2.5 Reconstruction of Simplified Genetic Programs

After the simplification process, we need to turn the simplified DDB form to the genetic program format and put them back to the population so that the evolutionary process can continue. This is done by tracking the record of the simplification process.

3 Experiment Design

3.1 Image Data Sets

In the experiment, we used four data sets providing object classification problems of varying difficulty. Example images are shown in Figure 3.

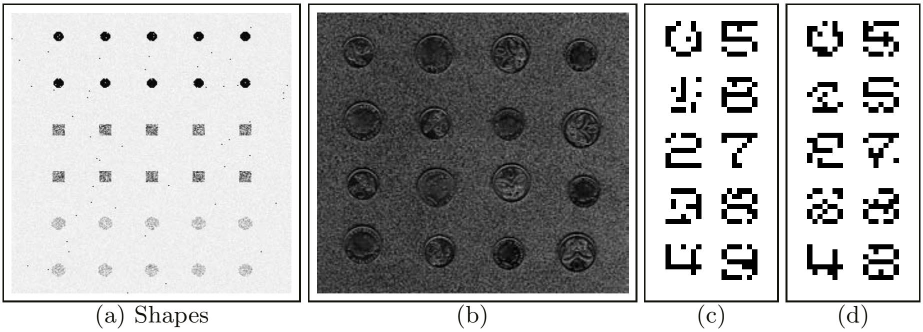

(a) Shapes (b) (c) (d)

Fig. 3. Sample Data sets. (a) Shape; (b) Coin; (c) Digits15; (d) Digits30

The first set of images (figure 3a) was generated to give well defined objects against a relatively clean background. The pixels of the objects were produced using a Gaussian generator with different means and variances for each class. Three classes of 960 small objects were cut out from those images to form the classification data set. The three classes are: black circles, grey squares, and light circles. For presentation convenience, this data set is referred to as *shape*.

The second set of images (figure 3b) contains scanned 5 cent and 10 cent New Zealand coins. The coins were located in different places with different orientations and appeared in different sides (head and tail). In addition, the background was cluttered. We need to distinguish different coins with different sides from the background. Five classes of 801 object cutouts were created: 160 5-cent heads, 160 5-cent tails, 160 10-cent heads, 160 10-cent tails, and the cluttered background (161 cutouts). Compared with the *shape* data set, the classification problem in this data set is much harder. Although these are still regular objects, the problem is very hard due to the noisy background and the low resolution.

The third and the fourth data sets are two digit recognition tasks, each consisting of 1000 digit examples. Each digit example is an image of 7×7 bitmap. In the two tasks, the goal is to automatically recognise which of the 10 classes (digits 0, 1, 2, ..., 9) each pattern (digit example) belongs to. Note that all the digit patterns have been corrupted by noise. In the two tasks (figures 3c and 3d), 15% and 30% of pixels, chosen at random, have been flipped. In data set 3, while some patterns can be clearly recognised by human eyes such as “0”, “2”, “5”, “7”, and possibly “4”, it is not easy to distinguish between “6”, “8”and “3”, even “1” and “5”. The task in data set 4 is even more difficult – human eyes cannot recognise majority of the patterns, particularly “8”, “9” and “3”, “5” and “6”, and even “1”, “2” and “0”. In addition, the number of classes is much greater than that in tasks 1 and 2, making the two tasks even more difficult.

For all the four data sets, the objects were equally split into three separate data sets: one third for the training set used directly for learning the genetic program classifiers, one third for the validation set for controlling overfitting, and one third for the test set for measuring the performance of the learned program classifiers.

3.2 Experiment Configuration

In the approach, we used the tree-structure to represent genetic programs [1]. The ramped half-and-half method was used for generating programs in the initial population and for the mutation operator [2]. The proportional selection mechanism and the reproduction, crossover and mutation operators [1] were used in the learning and evolutionary process.

In our configuration, four image features extracted from the objects were used as terminals first two tasks, and just 49 pixel values in the third and fourth tasks. In addition, we also used a constant terminal for these tasks. The function set consists of the four standard arithmetic operators and a conditional operator, $\{+, -, \times, \%, if\}$. We used classification accuracy on the training set of object images as the fitness function.

We used reproduction, mutation, and crossover rates of 10%, 30%, and 60%, respectively. The program depth was initialised from 3-6, and can be increased to 7 during evolution. The population size was 500. The evolutionary process was run for a maximum of 50 generations, unless it found a program that solved the problem perfectly (100% accuracy), at which point the evolution was terminated early. All single experiments were repeated 50 runs and the average results are presented in the next section.

4 Results and Discussion

Instead of simplifying genetic programs every generation during evolution, we applied the simplification algorithm to the evolutionary process every five generations on the four data sets and compared their performance with the basic GP approach without simplification. The average classification accuracy and the average training time over the 50 runs are presented in table 1.

Table 1. Object classification accuracy (in %) and the training time (in second)

Method	Classification accuracy (%)				Training time (second)			
	Shape	Coin	Digit15	Digit30	Shape	Coin	Digit15	Digit30
GP-basic	99.41	85.18	56.92	43.95	14.68	16.98	31.93	28.37
GP-Simplification	99.72	86.09	58.64	45.06	12.29	12.12	21.92	20.28

Classification Accuracy. According to table 1, the GP approach with the proposed simplification method always achieved better object classification accuracy on these data sets than the basic GP approach without simplification. The improvement over the basic GP approach is particularly clear for the two difficult digit data sets.

We hypothesised that the simplification process during evolution might destroy the existing good building blocks of the genetic programs, which might result in worse classification performance. However, these results are clearly different from the original hypothesis. After checking the evolutionary process, we found that this improvement was not clear at the beginning of evolution. At this stage, although the simplification algorithm might destroy some potential good building blocks, this effect was very much compensated by the powerful crossover operator, which can preserve good building blocks. At the later stage, when the GP evolution is difficult to make further improvement since the crossover operator starts to destroy good existing building blocks, the simplification algorithm actually produces new genetic materials which might contain new good building blocks by *restructuring* the entire genetic programs. This makes it possible to consider the simplification a new genetic operator in the future.

Efficiency. The training times of the evolutionary process on the four data sets are presented in table 1. As expected, the GP approach with the simplification greatly improved the training efficiency. It seems that the improvement is even bigger as the difficulty of the classification problems was increased. This is mainly because the simplification process removes the redundancy, makes the genetic programs shorter, and accordingly reduces the search space.

5 Conclusions

The goal of this paper was to develop an online program simplification approach in GP during the evolutionary process. This goal was successfully achieved by developing an algorithm for finding the greatest common sub-multisets in programs with the same family of operators, constructing the double decker bus structure for tidy-up rewriting of genetic programs, applying double hashing for tidy-up rewriting genetic programs with different family of operators and processing the conditional operator in the simplification process.

The new approach was examined and compared with the basic GP approach without simplification on four image classification problems of increasing difficulty. The results suggest that, the new simplification approach outperformed the basic GP approach in terms of both classification accuracy and training time on all of the four data sets. The more difficult the classification problems, the larger improvement in both classification accuracy and training time.

The online simplification during evolution seems to be able to reduce the search space. While it could introduce new genetic materials, it is not clear whether it destroy good building blocks in the early stage of evolution, which needs to be further investigated in the future.

The simplification algorithm described in the paper has a number of levels. In the future, it is very interesting to investigate whether the simplification on the genetic programs with the same family of operators only can improve the evolutionary efficiency and system effectiveness performance. We will also investigate what effects would be produced if we consider the simplification a new operator and put it into the function set.

Acknowledgement

We would like to thank Dr Peter Andreae for a number of useful discussions.

References

1. John R. Koza. *Genetic Programming II: Automatic Discovery of Reusable Programs*. Cambridge, Mass. : MIT Press, London, England, 1994.
2. Wolfgang Banzhaf, Peter Nordin, Robert E. Keller, and Frank D. Francone. *Genetic Programming: An Introduction on the Automatic Evolution of computer programs and its Applications*. Morgan Kaufmann Publishers, 1998.
3. J. Eggermont, A. E. Eiben, and J. I. van Hemert. A comparison of genetic programming variants for data classification. In *Proceedings of the Third Symposium on Intelligent Data Analysis (IDA-99), LNCS 1642*. Springer-Verlag, 1999.
4. Daniel Howard, Simon C. Roberts, and Conor Ryan. The boru data crawler for object detection tasks in machine vision. In *Applications of Evolutionary Computing*, volume 2279 of *LNCS*, pages 220–230. 2002. Springer-Verlag.
5. Riccardo Poli. Genetic programming for image analysis. In *Genetic Programming 1996: Proceedings of the First Annual Conference*, pages 363–368, Stanford University, CA, USA, 28–31 July 1996. MIT Press.
6. Andy Song, Thomas Loveard, and Victor Ciesielski:. Towards genetic programming for texture classification. In *Proceedings of the 14th Australian Joint Conference on Artificial Intelligence*, pages 461–472. Springer Verlag, 2001.
7. Thomas Loveard. Genetic programming with meta-search: Searching for a successful population within the classification domain. In *Proceedings of the Sixth European Conference on Genetic Programming*, volume 2610 of *LNCS*, pages 119–129. 2003. Springer Verlag.
8. Mengjie Zhang, Victor Ciesielski, and Peter Andreae. A domain independent window-approach to multiclass object detection using genetic programming. *EURASIP Journal on Signal Processing, Special Issue on Genetic and Evolutionary Computation for Signal Processing and Image Analysis*, 2003(8):841–859, 2003.
9. Will Smart and Mengjie Zhang. Classification strategies for image classification in genetic programming. In *Proceeding of Image and Vision Computing Conference*, pages 402–407, New Zealand, November 2003.
10. Walter Alden Tackett. Genetic programming for feature discovery and image discrimination. In *Proceedings of the 5th International Conference on Genetic Algorithms, ICGA-93*, pages 303–309, 1993. Morgan Kaufmann.
11. Mengjie Zhang and Victor Ciesielski. Genetic programming for multiple class object detection. In *Proceedings of the 12th Australian Joint Conference on Artificial Intelligence* , LNAI Volume 1747. pages 180–192. 1999. Springer-Verlag.
12. Joel Moses. Algebraic simplification a guide for the perplexed. *Proceedings of the second ACM symposium on Symbolic and algebraic manipulation*. Los Angeles, California, United States, 1971. Pages 282 - 304.
13. Yun Zhang. Genetic Programming for Multiple class Classification. *BSc (Honours) Research Project*. School of Mathematics, Statistics, and Computer Science, Victoria University of Wellington. 2004.
14. Will Smart. Genetic Programming for Multiclass Object Classification. *BSc (Honours) Research Project*. School of Mathematics, Statistics, and Computer Science, Victoria University of Wellington. 2003.

An Ontology-Supported Database Refurbishing Technique and Its Application in Mining GSM Trouble Shooting Rules

Bong-Horng Chu[1, 3], In-Kai Liao[2], and Cheng-Seen Ho[2, 4]

[1] Department of Electronic Engineering,
National Taiwan University of Science and Technology
43 Keelung Road Sec. 4, Taipei 106, Taiwan
ben@ailab2.et.ntust.edu.tw

[2] Department of Computer Science and Information Engineering,
National Taiwan University of Science and Technology
43 Keelung Road Sec. 4, Taipei 106, Taiwan
cheng-seen.ho@ieee.org

[3] Telecommunication Laboratories, Chunghwa Telecom Co.,
Ltd. 11 Lane 74 Hsin-Yi Road Sec. 4, Taipei, 106, Taiwan
benjamin@cht.com.tw

[4] Information School, Chung Kuo Institute of Technology,
56 Hsing-Lung Road Sec. 3, Taipei, 116, Taiwan
shawnho@mail.ckitc.edu.tw

Abstract. The real-life customer servicing databases share a useful characteristic which can be properly exploited to solve a common problem with the databases themselves. This useful characteristic is that either remark or memo fields are always included in the databases; customer service representatives can use these fields to write down specific things about the service records. This design helps alleviate the following common difficulty with the categorization of customer-service-related problems: customer requested service records are often *misclassified* owing to human ignorance or bad design of problem categorization. In this paper we propose an ontology-supported technique to preprocess the remark fields, trying to discover meaningful information to help recategorize misclassified service records. This process restores the database into one with more meaningful data in each record, which facilitates the mining of better association rules. The technique was applied to a real-life trouble shooting database obtained from a telecommunication company. The results show a substantial improvement in the quality of mined trouble shooting rules can be obtained.

1 Introduction

Customer servicing databases are one of the most used databases in any customer relationship management systems. A typical example is the GSM trouble shooting database, which contains trouble shooting records about the GSM (Global System for Mobile Communications) system for customers of a telecommunication company. To help statistical analysis, a database is usually designed to contain domain-specific categorizing fields which can take on proper values to reflect the physical semantics of a service record. Unfortunately, the predefined field values are hardly exhausted

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 997–1004, 2005.

and therefore values like "others" have to be introduced. As a consequence, the categorizing fields may contain scattered "others" and hinder subsequent analysis of the database. Fortunately, facing a "others" value, the representative usually also puts down something he/she comprehends about the service record into the "remark" fields so that other service personnel can figure out what information really lies behind later. To capture more information about a service record, a customer servicing database usually contains more than one remark field; some are used to describe customer's detailed intention, major complaints or requests, and others are used to keep track of service progress.

The good news is: we can introduce text mining techniques to retrieve significant information embedded in the remark fields to help discover the real meaning behind the "others" values. In this paper, we put this idea into practice. Taking a trouble shooting database of a GSM system as an example, we first preprocess the database to discover significant values from the remark fields, and then refurbish the database by replacing the "others" values with the discovered values. We also describe how ontology can support this process. Finally, we run a data mining module on the refurbished database to discover better trouble-shooting rules as a demonstration of the feasibility of our work.

The rest of the paper is organized as follows. Section 2 introduces our ontology-supported approach to database refurbishing. Section 3 discusses the data mining technique we used in discovering rules from the refurbished customer servicing database. In Section 4, we show the experimental results on how well our methods perform. We conclude in Section 5 by summarizing our contributions and discussing some further work.

2 Ontology-Supported Database Refurbishing

In our example application, the trouble shooting database for a GSM system contains three categorizing fields, one *symptom* field to record the customer's complaints, one *cause* field to note the causes behind the symptoms, and one *process* field to describe the processes the representatives take to resolve the causes, each field containing a set of categories, including the inevitable "others" category. In addition, there are two remark fields, one is associated with the *symptom* field, and the other is associated with both the *cause* and *process* fields.

Before performing data preprocessing, we followed the guidance of the construction procedures proposed by Noy and McGuinness [1] to develop three ontologies related to *symptoms*, *causes*, and *processes*, namely *GSMSO* (*GSM-trouble Symptom Ontology*), *GSMCO* (*GSM-trouble Cause Ontology*) and *GSMPO* (*GSM-trouble Process Ontology*), respectively, to define the conceptual terminologies and concept hierarchies within the GSM trouble shooting domain. Fig. 1 illustrates portions of the three ontologies. Taking GSMSO for explanation, it depicts a multi-level conceptual hierarchy of most possible GSM trouble symptoms. The first level contains several symptom categories, e.g. *connection abnormity*, *quality abnormity*, *setup abnormity*, *short message abnormity*, *handset abnormity*, etc. Under *handset abnormity* category, there are symptoms of *automatic shut down*, *automatic turn on*, etc. Possible synonyms are also associated with each symptom in the ontology, which are not shown in the figure. With the support of these ontologies, we can use the symptom remark to refine the *symptom* field, while using the other remark to refine the *cause* and *process* fields.

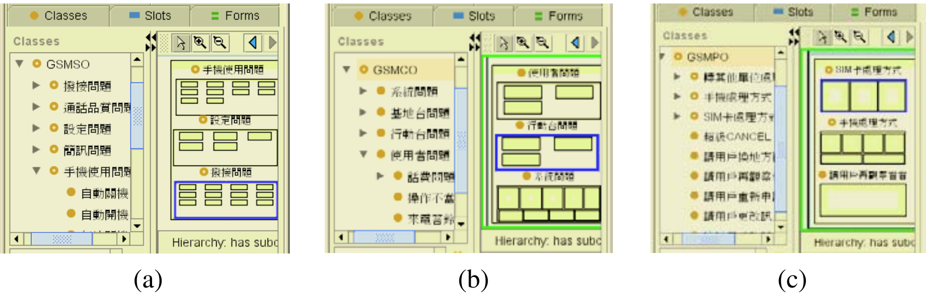


Fig. 1. Portions of three ontologies: (a) GSMSO; (b) GSMCO; (c) GSMPO. Note that all trouble shooting related terms are in Chinese in this research

Fig. 2 illustrates our method to do database refurbishing, which includes remarks grouping, keyword extraction, and terms refurbishing, to refine the *symptom* fields. The *cause* and *process* fields are treated similarly.

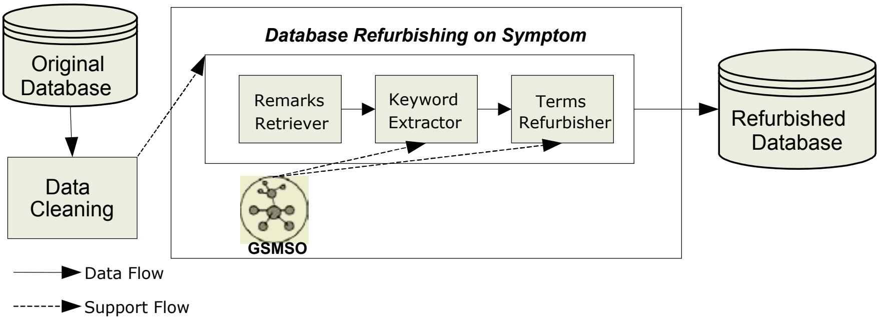


Fig. 2. Architecture for Database Refurbishing

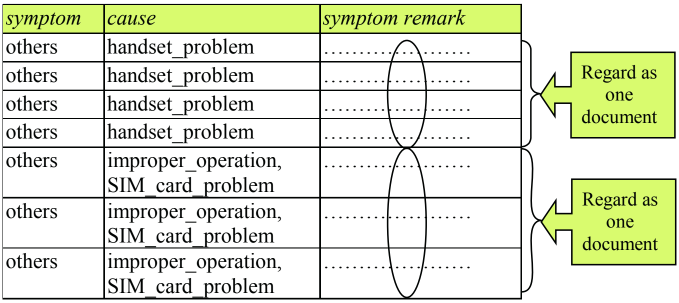
Remarks Grouping

The first module, Remarks Retriever, retrieves the symptom remarks from the records that have the value “others” in the *symptom* field, groups those remarks into different documents according to their *causes*, and then tokenizes the texts in each document. For illustration, Fig. 3 shows a document-grouping example.

Because the contents of remark fields in our example database are Chinese sentences, we have to tokenize those sentences, i.e., segmenting the sentences into Chinese terms, for subsequent process. We use MMSEG [2] to do tokenizing.

Keyword Extraction

After tokenizing, we found that many Chinese terms in the GSM trouble shooting domain cannot be discriminated very well by MMSEG. Keyword Extractor is thus introduced to fix the problem. First, supported by the *GSMSO* ontology, it fixes the wrong tokens segmented by MMSEG and re-organizes them into symptom terms contained in the ontology. It then eliminates the terms which cannot be found in the ontology. Only ontology terms or their synonyms can stay in the document after this step. Finally, all synonyms are transformed to canonical terms as used in the ontology.



symptom	*cause*	*symptom remark*
others	handset_problem	....
others	handset_problem	....
others	handset_problem	....
others	handset_problem	....
others	improper_operation, SIM_card_problem	....
others	improper_operation, SIM_card_problem	....
others	improper_operation, SIM_card_problem	....

Regard as one document
Regard as one document

Fig. 3. Remarks grouping

Terms Refurbishing

Now we need to identify significant terms inside each document as candidate categories for replacing category "others". We use the TFIDF (*term frequency – inverse document frequency*) technique proposed by Joachims [3] and Neto *et al.* [4]. Terms Refurbisher starts by utilizing TFIDF to calculate a weight for each term in the documents of symptom remarks, and then identify most significant terms with their weights over a pre-defined threshold. With these significant terms, Terms Refurbisher then can replace "others" with proper new terms in every record. The replacement has to depend on what significant terms are contained in the *symptom remark* field of a record. For example, suppose we have identified three significant terms, namely, A_term, B_term, and C_term. Terms Refurbisher can replace the term "others" in the *symptom* field with only "A_term" if the *symptom remark* field of the record contains only "A_term".

3 Data Mining Using PD with PH Algorithm

To discover association rules embedded in the exemplified database, we introduced perfect hashing (PH) to improve the PD algorithm [5] and developed a PD with PH algorithm. Note that the main defects of Apriori Algorithm are that it often produces a huge amount of candidate itemsets and needs to scan the whole database repeatedly during the mining process. Zou *et al.* have proposed a PD (Pattern Decomposition) technique trying to ameliorate them. PD decomposes transactions into short itemsets to make it possible for merging regular patterns together, thus reducing the dataset in each pass. PD consists of three cooperating processes, namely, *PD*, *PD-rebuild* and *PD-decompose*. The algorithm reduces the dataset whenever new infrequent itemsets are discovered. It comprises a set of passes starting from *pass 1* for the originally given dataset D_1. Each pass has two phases. First, frequent itemsets L_k and $\sim L_k$ are generated by counting a support for each k-itemset in D_k. Second, *PD-decompose* is used to decompose D_k to get D_{k+1} such that D_{k+1} contains no itemsets in $\sim L_k$. The algorithm terminates at a pass z if D_z is empty.

Fig. 4 shows how perfect hashing is incorporated into PD to form the PD with PH algorithm. At Step 6 we employ perfect hashing to remember all k-itemsets for each pattern p. This reduces processing time especially when the patterns are composite ones. Steps 8 and 9 introduce perfect hash table PH_k to speed the counting of supports for k-itemsets. Finally, at Step 12, we can get L_k and ~L_k directly from scanning perfect hash table PH_k. Details can be referenced from [6].

```
PD with PH ( transaction-set $T$ )
1:$D_1 = \{<t, 1>| t \in T \}$; $k$=1;
2:while ($D_k \neq \Phi$) do begin
3:  $PH_k$=an empty perfect hash table;
4:  forall $p$ in $D_k$ do begin
5:    $ph$=an empty perfect hash table;
      //get all k-itemset of p.IS with perfect hashing
6:    put all $k$-itemset $s \subseteq p.IS$ to $ph$
      //counting support with perfect hashing
7:    forall bucket $h \in ph$ do begin
8:      if $h$ in $PH_k$ then $PH_k.h.Occ$+=$p.Occ$;
9:        else put <$h$, $p.Occ$> to $PH_k$;
10:   end
11: end
12: decide $L_k$ and ~$L_k$ by scanning $PH_k$;
    //build $D_{k+1}$
13: $D_{k+1}$= PD- rebuild ($k$, $D_k$, $L_k$, ~$L_k$);
14: $k$++;
15:end
16:Answer = $\bigcup_k L_k$
```

Fig. 4. PD with PH

4 Empirical Evaluation

In this section, we describe how our ontology-supported database refurbishing method performs over the exemplified GSM trouble shooting database, which is a real-life trouble shooting historical database from one of the major telecommunication companies in Taiwan. To demonstrate the effectiveness of our technique, we will make a comparison about the results of association rules mined from both original and refurbished databases.

First, we performed data cleaning on 15022 data records by removing those records that contain ambiguous or irrelevant data values in the *symptom*, *cause*, or *process* fields from the perspective of trouble-shooting. For example, the "cannot_contact_customer" value in the *process* field would be removed because it provides no useful information for trouble-shooting. After data cleaning, we had 5234 data records left for experiments, 4188 records for training and 1046 records for testing. Finally we applied the PD with PH mining algorithm to the training dataset and obtain two sets of association rules, one in the form of *symptom*→*cause*[1] and the other *cause*→*process*. The *minimum support* was set to 1% and the *minimum confidence* was set to 40%.

Rules of *symptom*→*cause*

Without database refurbishing, we obtained seven rules of *symptom*→*cause* from mining the original training dataset, as shown in Table 1. Note that there are two rules which contain "others" in the *symptom* part, so only the rest of five rules are useful. The values in the "hit count" column of the table denote how many records in the training dataset match with the rules, so we can calculate the total accuracy rate by Equ. (1).

[1] *if-part*→*then-part* denotes that if the facts in the *if-part* occurred, then it should take the conclusions or actions in the *then-part*

$$\frac{summation\_of\_hit\_counts}{amount\_of\_training\_records} = \frac{1600}{4188} = 38.2\% \cdot \qquad (1)$$

Table 1. Rules of *symptom→cause* mined from the original training dataset

Rule id	Symptom	Cause	Hit count
Rule_1	dial_in_failure	handset_problem	472
rule_2	dial_out_failure	handset_problem	440
rule_3	noise_or_interference	handset_problem	232
rule_4	connect_failure	location_abnormality	191
rule_5	one_way_talk	handset_problem	265
rule_6	Others	sim_card_problem	N/A
rule_7	Others	improper_operation	N/A

Then we did three experiments with TFIDF weight threshold set to $\frac{tfidf_{\min} + tfidf_{\max}}{2} \times \frac{3}{2}$, $\frac{tfidf_{\min} + tfidf_{\max}}{2}$ and $\frac{tfidf_{\min} + tfidf_{\max}}{2} \times \frac{1}{2}$, respectively, where $tfidf_{\min}$ denotes the minimal TFIDF weight in the document, and $tfidf_{\max}$ the maximal one. The result shows that no matter what the setting is, it always can generate more and better rules than the original dataset. The largest total accuracy is obtained when the TFIDF weight threshold is set to $\frac{tfidf_{\min} + tfidf_{\max}}{2}$. Table 2 shows that ten more useful rules are mined from the refurbished training dataset with this weight threshold, compared with those in Table 1. Note that the antecedents of six out of ten new rules, i.e. rule_10 to rule_15, contain new terms introduced by our ontology-supported database refurbishing mechanism. The total accuracy of the rules in Table 2 is improved to 79.5%.

Table 2. More useful rules of *symptom→cause* mined from the refurbished training dataset

Rule id	Symptom	Cause	Hit count
rule_6	international_call_breakdown	improper_operation	106
rule_7	dial_in_failure	location_abnormality	190
rule_8	unknown_caller	system_problem	184
rule_9	connect_failure	sim_card_problem	250
rule_10	Sim_card_fault	sim_card_problem	241
rule_11	vms_password_reset_failure	vms_password_reset	175
rule_12	basic_rbt_set_failure	system_problem	165
rule_13	receive_only_failure	improper_operation	154
rule_14	weak_signal	base_station_problem	166
rule_15	sms_transmit_breakdown	handset_problem	99

Rules of *cause→process*

Similar to the above process, we discovered six rules of *cause→process* from the original training dataset. Only three of them are useful, however. The total accuracy is 33.4%. We also did three experiments with different TFIDF weight thresholds as above, and found the latter two have the same rules mined with the same total accuracy, both outperforming the first one. The total accuracy of them is raised to 72.5%.

After making comparisons of accuracy between the rules mined from the original and refurbished training datasets, we found the best results appeared when the TFIDF weight threshold was set to $\frac{tfidf_{\min} + tfidf_{\max}}{2}$. We therefore accepted it in the final model and evaluated the model using the testing dataset of 1046 records. The results of the evaluation are shown in Table 3.

Table 3. Evaluation of the relatively best model on the testing dataset

	symptom→cause	*cause→process*
	Total accuracy	Total accuracy
$TFIDF\_threshold = \frac{tfidf_{\min} + tfidf_{\max}}{2}$	72.4%	61%

5 Conclusions

We have described an ontology-supported database refurbishing technique to refurbish customer servicing databases so that better results can be obtained in subsequent data analysis or data mining. Supported by ontology, the technique can successfully identify most significant terms from a group of remarks; remarks are collected into a group according to some proper category field. The significant terms are then used to refurbish the categorizing fields by replacing those useless "others" values with meaningful terms. To demonstrate the effectiveness of our approach, an exemplified real-life GSM system trouble shooting database are refurbished, with the support of three domain-related ontologies. The refurbished datasets are then submitted for data mining by an enhanced PD with PH algorithm to discover trouble shooting rules embedded in the database. Our experiments show the technique allows more significant terms and more useful trouble shooting rules to be derived from the refurbished database and the total accuracy of the set of mined rules are much better than that from the original database.

The total accuracy of one set of rules on the testing dataset is only over 61%. This fact has attracted us to search for further improvement. One way of improvement is to apply the refurbishing technique to every categorizing field, rather than only remark fields, to produce a better refurbished database. The second way is to find an alternative data mining technique which is more suitable to the exemplified database than association rule mining. Both are under our investigation.

References

1. Noy, N.F., McGuinness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March (2001)
2. Tsai, C.H.: MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of The Maximum Matching Alforithm. Availible at http://www.geocities.com/hao510/mmseg/ (1996)
3. Joachims, T.: A Probabilistic Analysis of the Rocchio Algrithm with TFIDF for Text Careagorization. Technical Report of CMU-CS-96-118, Department of Computer Science, Carnegie Mellon University, Pennsylvania, USA, March (1996)

4. Neto, J.L., Santos, A., Kaestner, C., Freitas, A.: Document Clustering and Text Summarization. In: Proceedings of the Fourth International Conference Practical Applications of Knowledge Discovery and Data Mining (PADD-2000), London, January (2000) 41-55
5. Zou, Q., Chu, W., David, J., Chiu, H.: A Pattern Decomposition Algorithm for Data Mining of Frequent Patterns. Knowledge and Information Systems 4 (2002) 466-482
6. Liao, B.C.: An Intelligent Proxy Agent for FAQ Service. Master Thesis, Department of Electronic Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C. (2003)

Develop Secure Database System with Security Extended ER Model*

Xin Liu[1,2], Zhen Han[1], Jiqiang Liu[1], and Chang-xiang Shen[1,3]

[1] Research Center of Information Security Architecture, Beijing Jiaotong University,
Beijing, China
zhan@center.njtu.edu.cn
[2] Department 4, Naval University of Engineering, Wuhan, China
jfz97@sohu.com
[3] The Naval Compute Technology Institute, Beijing, China

Abstract. Database system security is desired for acutely today. But it is difficult and expensive to develop secure database system because the developers usually find it difficult to design a consistent and complete database structure. To solve the problem we focus on the research results on both DB design model and the security mechanisms. The secure database system development will be simplified greatly by combining database design methodology and information security technique. So we made the effort to combine the security control mechanisms (e.g. MAC) and database develop model—Entity Relation (ER) model to construct a Security Extended ER (SEER) model. The security of the database system developed with SEER model is analyzed at the last part in the paper. The research and applying of SEER model puts forward a new ideal about secure database system development.

1 Introduction

Developing secure database system is an expensive and time-consuming task. With the increasing severity of the security problems of the database systems, many research results were reported in the field of secure database management system while some papers introduced the explore on the database system design method. However there is no report about how to apply the database developing methodology to the secure database system developing process up to date. The most popular secure DBMS is multilevel relational DBMS satisfying the class B1 or above criteria of TCSEC [1, 2]. But the design process exploring indicates that to design and develop secure database is neither simple nor easy even though the secure DBMS is used to develop database system. The increasing popularity of database desires guarantee of their security acutely while the developers usually find it difficult to articulate a consistent and complete set of security requirements. The Entity relation model is a powerful tool for database design. So the motivation of the paper is to combine the security control mechanisms and database develop methodology by applying information security technique to the Entity-Relation model (ER model) [3] to construct a security extended ER (SEER) model.

* The paper is supported by 973 project of China (Grant No. TG1999035801)

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1005–1010, 2005.

1.1 Database Security Design Overview

The database security involves five phases including security requirements analysis, security policies establishment, security abstraction, security implementation and security management. The nodes that are vulnerable to attacks corresponding storage security, transportation and access security. The security mechanisms include Identity and Authority, MAC, DAC, RBAC, Intrusion detection and audit etc.

1.2 The Thirst for Secure Database System Design Methodology

Examples of database design by two different Secure DBMS—SeaView [4] and Sword [5] are given in [6]. It is made clear that the design is not an easy issue and even a master in database can hardly finish it without secure knowledge by a simple database implementation example. So the methodology and toolkits of database security design are thirsted for (See [6] for details).

The organization of this paper is as follows: The theoretical basis of the project including the information security methodology and database design methodology is discussed in section two. The combination of the two techniques is illustrated at the end of this section. In section 3 the security of database developed with SEER model is analyzed followed by the conclusion in section 4.

2 Methodology and Model

2.1 ER Model

We focus on the common database design model—Entity Relation model first in order to find a way of solving the security design. The entity relationship approach is widely accepted as a useful basis for information modeling. The ER model adopts the view that the real world consists of entities and relationships among them. An entity is a ‘thing’ that can be distinctly identified. A relationship is an association among entities.

2.2 Mandatory Access Control

We explore the security extension on the ER model and only MAC mechanism is explained in details here. Mandatory access control is an important symbol of class B2 and above security system of TESEC. It is based on BLP model [7]. There are four components in the model: subjects (active entities such as users, processes), objects (passive entities such as files, data), access mode (read only, append, write, execute) and security. System defines some sensitivity levels in a partial order. For an example, the levels are Unclassified, Confidential, Secrete and Top Secrete in an order U<C<S<TS. The security level of a subject reflects the authorization of the subject to the information. The maximum level is its security clearance. The security level of an object is its security class that reflects the protection requirement to that object. The access properties are summarized as “No-Up-read” and “No-Down-Write” rules.

2.3 SEER Model

We extend the ER model by introducing an entity attribution named security attribution and a relationship named security relationship. Security attribution records security class of entity. Security relationship presents the access mode of a subject to an object. Security attribution satisfies a partial relation and the relation between a subject and an object decides whether the subject can access the object or not in a specific access mode. We call the modified model as security extended entity relationship model (SEER). The finest granularity of the security attribution is tuple. It is well illustrated in figure1.

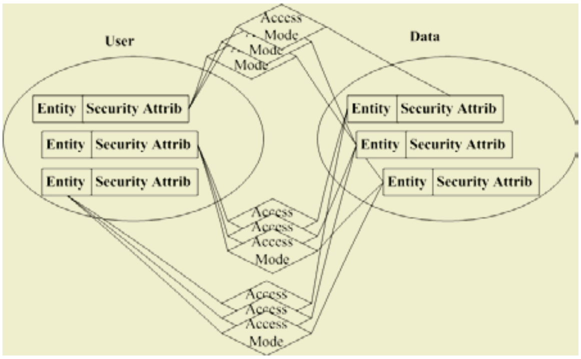


Fig. 1. SEER Model Sketch

Security attribution works as security label in database system.

The SEER model can be used to realize the mandatory access control mechanism in database systems. A security attribution presents clearance or class of an entity and a security relation presents the access mode (read only, write only, read-write) the subject can do to the object. The Access mode is gained by calculating the domination relationship between the security attributions of entities.

3 Security of Database System Developed with SEER Model

3.1 Overview

We use a security enhancement Linux as the OS and a source open software PostgreSQL [8] as the DBMS in the database server. The purpose of using a trusted OS is to apply the security control mechanism of OS well to ease the database security design burden. OS is responsible for the startup of DBMS and the access control of the database files.

The developers specify the database organization and define the meta data of the target system with SEER model. So the work of security define is standard, unified and quick. The security functions construct the “Security Module” block in Figure 2. The difficult part is the work of giving the PostgreSQL the ability to deal with MLS relations and SEER model is used to deal with the problem.

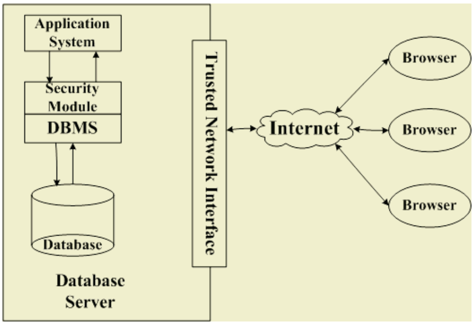


Fig. 2. The secure database system architecture

3.2 MAC Mechanism

Mandatory access control is an important characteristic of security system above class B1 and powerful means to avoid Trojan attack. SEER model guarantees that data in the database system can only be accessed by authorized subjects via correct access mode thus MAC function is realized in database system. A label module is needed to realize the entity security attribution and security relation embodied as follows: (1) to provide interface of label definition to define the concrete value of security class of database system for example top secrete, secrete or category names. In the developing process of database system, the definitions can be modified but this function is disabled for ordinary users after the system is applied. (2) to provide label management interface including the initiation and modification of clearances of subjects or classes of objects. The labels of users are stored in user-clearance table and the labels of the objects are stored in obj-class table. They are two system tables.

During the running period of database system, the security module receives sentences from client then call relative security check function to judge whether the sentences are satisfied with the security rules according to the label information along with other security parameters for DAC, RBAC. The sentences are handed to DBMS if they are satisfied or are denied and returned warning information otherwise. The result tuples whose classes are higher than those of the subject is cancelled before they are returned to the users in query.

3.3 Security Analysis of the Database System

Identification & Authorization in database system is based on the trust on Kerberos, DAC and RBAC is based on the intrinsic mechanism of PostgreSQL so they are trusty. The security of Convert channel analysis and MAC modules need to be verified and the later will be discussed here. The OS is trusted in our architecture and the MAC is desired to be realized on finer granularity—tuple. Will it be effective?

Seen from figure 2 the security modules separate application and DBMS. With the guarantee of security operation system, no program can access the DBMS without the security checking from security module directly.

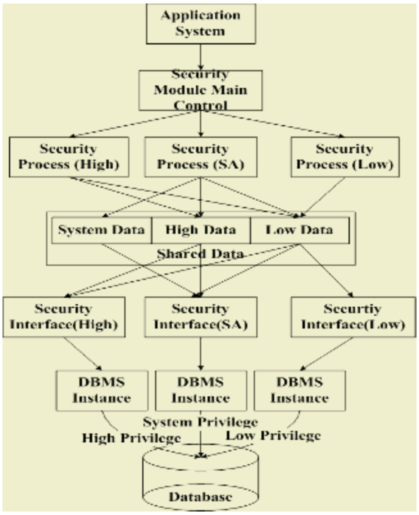


Fig. 3. Security Process Architecture

Whenever a browser's request is sent to the server program of the application, the security module will startup a security process in the user's level to deal with it. The architecture is depicted to show how it works. The security process capture each SQL command from application process and check them by access rules.

Every security process and security interface process has security class decided by the main control process after identification. The class is controlled by the OS, so the shared memory, disk buffers are restricted by MAC of OS. The subject class is controlled.

So as to objects, PostgreSQL stores physical files in the unit of relation. The objects at relation granularity are protected by OS because every relation has its label of OS.

Every tuple has its class attribution. If the class is higher than the subject's class, the main control process will create a temp process having the same class with the tuple. The temp process computes the query result in the corresponding memory and writes the result in lower class into lower memory. So the high tuple won't be in lower memory and won't be access by lower process.

It is proven that the system developed using the toolkit will realize the MAC mechanism through the above illustration and other security mechanisms are realized also.

4 Conclusions

We proposed the security extended entity relation (SEER) model by combining the database design theory and information security theory to provide a useful guideline to database system security design. The database system developed with the SEER model is able to implement the MAC mechanism on the granularity of tuple. The research and develop of SEER model gives a new idea of secure database system development.

References

1. Department of Defense: Trusted Computer System Evaluation Criteria, DoD 5200.28-STD, December (1985).
2. NCSC: The Trusted Database Management System Interpretation of Trusted Computer System Evaluation Criteria, NCSC-TG-21, April, (1991)
3. Chen PP, The Entity-Relationship Model: Towards & Unified View of Data, ACM Transaction on Database Systems, Vol 1, No.1, (1976), pp9-36.
4. Lunt, T.F.; Denning, D.E.; Schell, R.R.; Heckman, M.; Shockley, W.R, The SeaView security model, IEEE Transactions on Software Engineering, Vol 16, Issue: 6 , June (1990), pp593 - 607
5. A.W.Wood, S.R.Lewis, S.R.Wiseman, The SWORD Multilevel Secure DBMS, DRA Report No. 92005,(1992)
6. Lewis, S.; Wiseman, S.; Database design and MLS DBMSs: an unhappy alliance? Computer Security Applications Conference, 1993. Proceedings., Ninth Annual , 6-10 Dec. (1993), pp232 – 243
7. Bell, D.E., and La Padula, L.J. Secure Computer Systems: A Refinement of the Mathematical Fundations [J], ESD-TR-73-278, Vol.III, AD 780 528, Electronic Systems Division, Air Force System Command, Hanscom AFB, Bedford, Massachusetts, November. (1974)
8. The PostgreSQL Global Development Group, PostgreSQL 7.3.2 Developer's Guide [S/OL], http://www.postgresql.org/docs/pdf/7.3/developer-7.3.2-A4.pdf，2004/4-25

An Inference Detection Algorithm Based on Related Tuples Mining

Binge Cui and Daxin Liu

Computer Science and Technology Institute, Harbin Engineering University, Harbin
150001 Heilongjiang, China
cuibinge@yahoo.com.cn

Abstract. Existing algorithms on inference detection for database systems mainly employ functional dependencies in the database schema to detect inference, but what they can detect is limited. This paper presents a new data level inference detection algorithm. It can determine whether sensitive information can be disclosed from the user's query history through finding the related tuples between the return results of different queries. If two tuples are related to each other, then they will be merged into one tuple, thus the query history can be compressed. Moreover, the merged tuple has more information than the original two or more tuples. The experiment results show that, as the query number increases, our algorithm can infer almost the whole original relation; meanwhile the query history is compressed remarkablely. The system administrator should restrict user's query number and category to ensure that the database is secure.

1 Introduction

In a multilevel database system an inference channel occurs if users are able to infer some data that they are not allowed to access [1]. Inference channels allow users to obtain sensitive information without directly accessing them, thus it subverts the access control in database systems. Techniques to detect and remove inference channels can be categorized into two categories: The first category includes techniques that detect inference channels during database design; any channels are removed by modifying the database design and / or by increasing the classification levels of some of the data items [2], [3], [4]. These techniques often can result in overclassification of data and, therefore, reduce the availability of data. Moreover, inference detection during database design can't prevent inference channels that are introduced during query time.

Techniques in the second category seek to detect inference channel during query time [5], [6], [7], which include inference based on the database constraints and based on the data itself. The disclosure inference algorithms proposed by Brodsky [5] belong to the former case, which can generate all the information that can be disclosed to a user based on the user's past and present queries and the database and metadata constraints. However, they consider only the selection condition of the query a conjunction of equalities, but not a conjunction of equalities and inequalities. Thus the information included in the inequalities of the selection condition is lost, and the inference result is incomplete. The algorithms proposed by Yip and Levitt [6], [7] are

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1011–1017, 2005.

based on the analysis of the data stored in the database. They have identified six inference rules that a user can use to perform inferences. Inference detection during query processing has a common shortcoming, that is storing and processing a user's query history need to consume a lot of system resources every time.

In this paper, we describe our effort in developing a data level inference detection algorithm. It compares the user's present query with previous issued queries to find the related tuples between the return results of the queries. When two tuples are found to relate to each other, they will be merged into one tuple. If sensitive information is found in the new merged tuple, then the user's query will be rejected. Because all related tuples are merged into one tuple, our algorithm can compress the user's query history. The experiment results show that, as the query number increases, our algorithm can infer almost the whole original relation, meanwhile the query history can be compressed remarkably.

2 Inference Rules

We consider inference detection in a relational database with a single table. Ai denotes an attribute in the table, and ai denotes an attribute value from the domain of Ai. t[Ai] denotes the attribute value of a single tuple t over the attribute Ai. A query is represented by a 2-tuple (attribute-set, selection- criterion), where attribute-set is the set of attributes projected by the query, and selection-criterion is the logical expression that is satisfied each return tuple of the query. In general, Qi refers to the query (ASi, SCi), and {Qi} denotes the set of return tuples of Qi. 'Ç', 'È', and '\' stand for the set intersection, union, and difference operation respectively. Now, we introduce several notions that are used throughout this paper.

Definition 1. A tuple t over a set of attributes AS 'Satisfy' a logical expression E if E is evaluated to true when each occurrence of A_i in E is instantiated with $t[A_i]$, for all A_i in AS. t 'contradicts' with E if E is evaluated to false.

Definition 2. Given two queries, Q_1 and Q_2, we say that Q_1 is 'subsumed' by Q_2, denotes as $Q_1 \subset Q_2$, iff

$SC_1 \rightarrow SC_2$; or

for each tuple t_1 in $\{Q_1\}$, t_1 satisfies SC_2.

where '→' is the logical implication. '⊂' is a reflexive, anti-symmetric, and transitive relation.

Definition 3. A return tuple t_1 'relate' to another return tuple t_2 if the two tuples are selected from the same tuple in the database. Hence, $Q_1 \subset Q_2$ implies that for each return tuple t_1 of Q_1, there is a return tuple t_2 of Q_2, such that t_1 relates to t_2.

Given the above basic notations, we will describe the four inference rules. These rules can be used to find and merge the related tuples between the return results of different queries.

Inference Rule 1 (Subsume) Given two queries Q_1 and Q_2, such that $Q_1 \subset Q_2$, SI1 If there is a attribute A in $(AS_2 \setminus AS_1)$, such that all return tuples of Q_2 take the same attribute value a over A, then for each return tuple t_1 of Q_1, $t_1[A] = a$.

SI2 If t_1 is a return tuple of Q_1, t_2 is the only return tuple of Q_2 that is indistinguishable from t_1 (i.e. for all A in $AS_2 \cap AS_1$, $t_2[A] = t_1[A]$), then t_1 relates to t_2.

$Q_1 \subset Q_2$ implies that for each return tuple t_1 of Q_1, there is a return tuple t_2 of Q_2, such that t_1 relates to t_2. Thus if there is only one return tuple t_2 of Q_2 that is indistinguishable from t_1, then t_1 relates to t_2.

Inference Rule 2 (Belonging to) Given two queries Q_1 and Q_2, if t_1 is a return tuple of Q_1 such that t_1 satisfies SC_2, and t_2 is the only return tuple of Q_2 that is indistinguishable from t_1 (i.e. for all A in $AS_2 \cap AS_1$, $t_2[A] = t_1[A]$), then t_1 relates to t_2.

t_1 satisfies SC_2 implies that there is a return tuple t_2 of Q_2 such that t_1 relates to t_2. Thus if there is only one return tuple t_2 of Q_2 that is indistinguishable from t_1, then t_1 relates to t_2.

Inference Rule 3 (Overlapping) Given three queries Q_1, Q_2 and Q_3, if there is a return tuple t_1 of Q_1 that relates to a return tuple t_2 of Q_2, and t_1 also relates to a return tuple t_3 of Q_3, then t_2 relates to t_3.

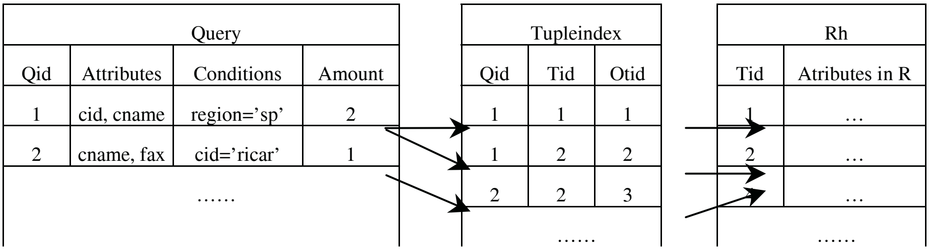
t_1 and t_2 are selected from the same tuple in the database, t_1 and t_3 are selected from the same tuple in the database too, thus t_2 and t_3 must be selected from the same tuple in the database.

Inference Rule 4 (Merge) Given two queries Q_1 and Q_2, if there is a return tuple t_1 of Q_1 that relates to a return tuple t_2 of Q_2, then t_1 and t_2 can be merged into one tuple t. t is defined as follows: for every $A_i \in AS_1 \cup AS_2$, $t[A_i] = \begin{cases} t_1[A_i] & A_i \in AS_1 \\ t_2[A_i] & A_i \in AS_2 \setminus AS_1 \end{cases}$

By using the Merge rule, all related tuples can be merged into one tuple that has more information than that of the original tuples [8]. Queries can adjust their pointers to point to the new tuple, thus Merge operation doesn't affect the user's access to query history. On the contrary, Merge operation can save a lot of storage space for users.

3 Inference Detection Algorithm

In this section, we will describe the inference detection algorithm in detail. First, we need to create three tables, Query, Tupleindex and Rh to store the user's query history. The relation schema for table Query is: Query (Qid, Attributes, Conditions, Amount), where Qid denotes the query number. The relation schema for table Rh is the same as that for table R, except that it has an additional attribute Tid, which is used to record the tuple number. Table Tupleindex is used to associate query with it's return tuples, whose relation schema is: Tupleindex(Qid, Tid, Otid). Qtid and Tid point to the same tuple in table Rh at the beginning, but Tid may point to another tuple later, due to the merge of related tuples. However, Qtid remain unchanged forever, and it will be used in the following Merge function. An example of the relations among the three tables is given in Figure 1. As it was shown, in table Rh the tuples whose Otid are 2 and 3 respectively are merged into one tuple, thus the user's query history can be compressed.



Query			
Qid	Attributes	Conditions	Amount
1	cid, cname	region='sp'	2
2	cname, fax	cid='ricar'	1
……			

Tupleindex		
Qid	Tid	Otid
1	1	1
1	2	2
2	2	3
……		

Rh	
Tid	Atributes in R
1	…
2	…
4	…
……	

Fig.1. Three Tables for Storing Query History

Our inference detection algorithm will be called each time a user U issues a query Q_i to the database, which is defined as follows.

```
INFERENCE(U, $Q_i$)
Insert $Q_i$ into table Query;
Insert all the return tuples of $Q_i$ into table Rh;
Insert the index record (Tid) of every return tuple of $Q_i$ into table Tupleindex;
Retrieve all queries from the table Query that is issued by user U (excluding $Q_i$). Let Q denote the query set;
for each $Q_j \in Q$ do
  for each $t_i \in Q_i$ do
    if Subsume($Q_j$, $Q_i$) or Satisfy($t_i$, $SC_j$) then
      if there is only one tuple $t_j \in Q_j$ such that $t_i[A]=t_j[A]$ ($\forall A \in AS_1 \cap AS_2$) then
        Merge($AS_i$, $t_i$, $AS_j$, $t_j$);
  for each $t_j \in Q_j$ do
    if Subsume($Q_i$, $Q_j$) or Satisfy($t_j$, $SC_i$) then
      if there is only one tuple $t_i \in Q_i$ such that $t_j[A]=t_i[A]$ ($\forall A \in AS_1 \cap AS_2$) then
        Merge($AS_i$, $t_i$, $AS_j$, $t_j$);
FIND_SENSITIVE_DATA(Rh);
```

The function Merge(AS_i, t_i, AS_j, t_j) is used to merge two related tuples, which is defined as follows.

```
Merge($AS_i$, $t_i$, $AS_j$, $t_j$)
if $t_i$[Tid]=$t_i$[Otid] then   '$t_i$ hasn't been merged into other tuples
  $t_j[A]=t_i[A]$ ($\forall A \in AS_i \backslash AS_j$)      'Merge $t_i$ into $t_j$
  delete $t_i$
  update $t_i$[Tid]=$t_j$[Tid]
else   '$t_i$ has been merged into other tuples
  $t_i[A]=t_j[A]$ ($\forall A \in AS_j \backslash AS_i$)      'Merge $t_j$ into the tuple which $t_i$ has been merged into before
  delete $t_j$
  update $t_j$[Tid]=$t_i$[Tid];
commit;
```

The function FIND_SENSITIVE_DATA(Rh) is used to find all possible inferences. If any sensitive information is found in table Rh, then the user's query will be rejected. The types of sensitive information vary with different system definition, but most of them can be found through the filter of query history.

4 Experimental Results and Analysis

The purpose of our experiments is to verify the efficiency of our inference detection algorithm, i.e., we want to know that if more information can be inferred from the user's query history as well as if the query history can be compressed effectively. The table used in our experiments is table customers in the sample database Northwind provided by Microsoft SQL Server 2000. This table has 91 records, and every record has 11 attribute values. There is no other functional dependency in this table except primary key dependency. The programming language adopted in our experiments is PowerBuilder 8.0, and the operating system is Windows 2000 Server. The queries are generated randomly, and every query comprises 5 fields. We assume the association between Customerid and Contactname is sensitive, that is to say, all queries can't contain the two attributes simultaneously. The experiment results are shown in figure 2 to figure 4.

Experiment 1 investigates the effect of the number of query on the average tuple restoration ratio. Here, we define the restoration ratio as the ratio of number of attribute value of every tuple in the query history to the number of attribute value of its' corresponding tuple in the original relation. It is obvious that the higher the restoration ratio, the more the similarity between the tuple in the query history and its' corresponding tuple in the original relation, and the more the risk that sensitive data being disclosed. From figure 2 we can see that, when the query number reaches 20, the average restoration ratio of every tuple has reached 98%. That is to say, most tuples of the original relation has been completely disclosed except a few. Thus it will be dangerous if we don't restrict the number and category of user's queries.

Experiment 2 investigates the effect of the inference algorithm on the magnitude of query history. The two lines in figure 3 correspond to the cases where the query his-

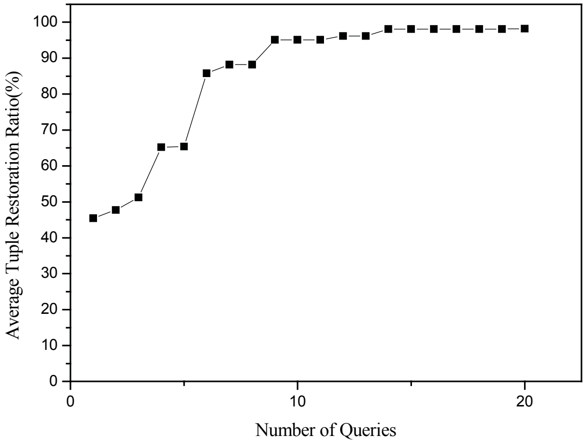


Fig. 2. Effect of the number of query on the average tuple restoration ratio

tory is compressed and not compressed respectively. It can be seen that, as the query number increases, the magnitude of query history grows rapidly when not compressed, and the cost of storing and processing the query history will be very high. However, when the query history is compressed, the tuple number will grow slowly at the beginning, and then decrease evenly, at last it will reduce to the size of the original relation. The system performance will also become better due to the decrease of system overhead.

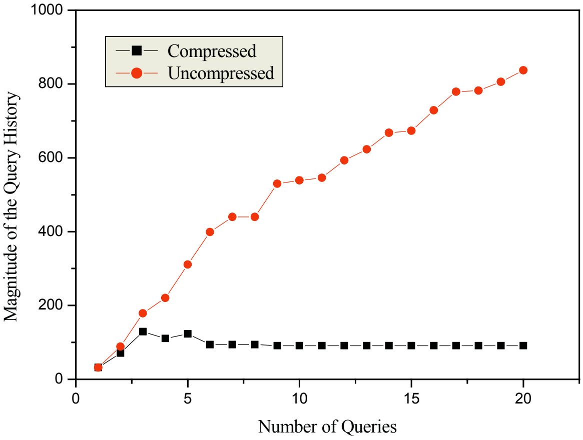


Fig. 3. Effect of the number of queries on the magnitude of query history

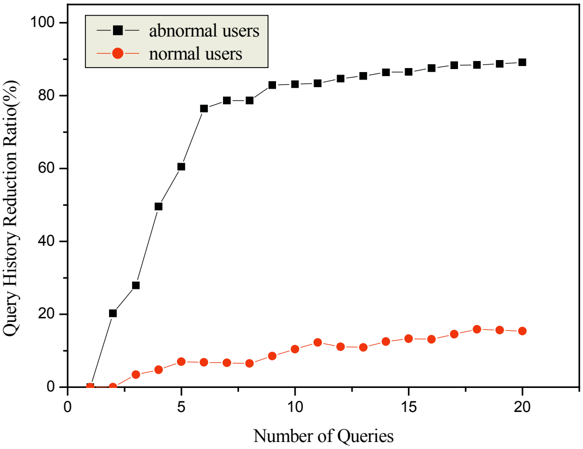


Fig. 4. Effect of the number of queries on the query history reduction ratio

Figure 4 also reflects the effect of the inference algorithm on the query history. In this figure, the dark line grows rapidly, which corresponds to the query history reduction ratio of those users who issue many queries and every query returns a lot of tuples. However, the red line grows slowly, which corresponds to those users who issue some queries randomly. The reason is that there are a lot of related tuples in the former case, thus the query history is compressed greatly. However, in the latter case the number of related tuples is relative small and thus the query history is compressed only a little. It is obvious that users aren't intended to inference from the query history in the latter case, thus the query history for them can be removed from the database. This approach can save the storage and computational costs for the database system.

From the above two experiments we can see that, by using the inference detection algorithm that is proposed in this paper, any user can infer almost all the tuples of the original relation from it's query history, though we can set some limitations to the user's behavior (e.g. some attributes of a relation are not allowed to present in the same query). Therefore, a system security administrator must restrict the user's query number and category. Furthermore, our data level inference detection algorithm should be used to determine whether there is any sensitive information disclosed from the user's query history.

5 Conclusions

In this paper, we proposed a new data level inference detection algorithm. We give four inference rules to find and merge the related tuples between the return results of different queries. Our algorithm can infer almost all the tuples of the original relation from the user's query history, thus it is necessary to restrict the user's query number and category. In other words, if any sensitive information can be inferred from the user's query history, then the new query of that user should be rejected. Our algorithm can also compress the user's query history greatly at the same time. This type of approach can be used as a vulnerability assessment tool of the database system, or as a tool for the retrospective analysis of log files after a suspected violation of confidentiality.

References

1. Farkas, C., Jajodia, S.: The Inference Problem: A Survey. ACM SIGKDD Explorations Newsletter, Vol. 4, No. 2 (2002) 6-11
2. Tsai, C.R., Gligor, V.D.: Inference Aggregation Detection In Database Management Systems. Proc. Of the IEEE Symposium on Research in Security and Privacy (1988) 96-106
3. Morgenstern, M.: Security And Inference In Multilevel Database And Knowledge-Base Systems. Proc. of The ACM SIGMOD International Conference on Management of Data (1987) 357-373
4. Su, T.A., Ozsoyoglu, G.: Controlling FD and MVD Inferences in Multilevel Relational Database Systems. IEEE Transactions on Knowledge and Data Engineering, Vol.3 (1991) 474-485
5. Brodsky, A., Farkas, C., Jojodia, S.: Secure Databases: Constraints, Inference Channels, and Monitoring Disclosures. IEEE Transactions on Knowledge and Data Engineering, Vol. 12 (2000) 900-919
6. Yip R., Levitt, K.: Data Level Inference Detection in Database Systems. Proc. of the 11th IEEE Computer Security Foundations Workshop (1998) 179-189
7. Yip R., Levitt, K.: The Design and Implementation of A Data Level Database Inference Detection System Proc. of the 12th Annual IFIP WG 11.3 Working Conference on Database Security. Chalkidiki, Greece (1998)
8. Qian, X., Lunt, T.F.: Tuple-level vs. Element-level Classification. Database Security, VI: status and prospects, Vancouver, Canada (1993) 301-315

A Preliminary Design for a Privacy-Friendly Free P2P Media File Distribution System

Ron G. van Schyndel

School of Computer Science and Information Technology
RMIT University PO Box 2476v, Melbourne 3000, Victoria
ron.vanschyndel@ieee.org,
http://www.cs.rmit.edu.au/~ronvs/

Abstract. In most P2P business models, in which users purchase the media, it is necessary to securely identify the user in order to facilitate payment. This paper presents a technique for allowing the widespread sharing of certain media formats including music using a method that keeps track of media possession and other marketing information, but in a way that does not require user identification. For the user, the main attraction of this scheme is that their identity is not a requirement, usage of reduced-quality media within this system is free and that extended media search is facilitated as an attraction to remain within the system. The content creators and distributors are compensated by this system by them having access to potentially large-scale actual usage and music trading statistics. The preliminary system design presented here, can cleanly coexist with a full-quality music purchase business model, also described briefly.

1 Introduction

In recent years, peer-to-peer networking has become an efficient and particularly scalable mechanism for widely distributing large media and data files. One application class that has proliferated on P2P networks is file-sharing and in particular, media-file sharing. This has been largely due to the popularisation of programs such as the original Napster [16], KaZaa [12] and Altnet [2], Gnutella [8], eDonkey [6] and many others. However, until 2003, most of this file sharing activity has been in breach of copyright law.

In 2003 Apple introduced iTunes [4] – a DRM controlled legal music-sharing system that allowed music-sharing under certain strict conditions. Since then, many other systems have emerged (Walmart [18], Musicmatch [15] – with Microsoft being the most recent to date [14] – each with minor variations to the same set of conditions).

All the above legal schemes have a subscription or pay-per-song business model. This model forces the media objects to assume the status of a commodity – a user must buy the object. This necessarily requires the user to provide their identity so as to secure payment.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1018–1024, 2005.

This paper describes, in an overview form, an initial design for a possible mechanism that entirely bypasses the need for user identity in securing access to music files available within this scheme.

An approach is presented where the economic value is not so much embedded in the shared media objects themselves, but in their relationship with each other – their context. While it is relatively easy to 'hijack' a media object such as a song which has intrinsic value and then pirate its value, it may not be so easy to hijack context.

This view directly benefits music publishers, which can use the context information for marketing. This marketing information is highly sought-after, especially if it is demonstrably representative of large population segments.

The information is not collected directly from a user's PC, but from the ISP through which the content is delivered.

1.1 A User-Centred Approach

The design of this system commences from the user's view of the media in a business model which is a little more involved than a simple fee-for-service or fee-for-product.

In this system, a user collects media tags and uses them to facilitate media playing and exchange. A typical media-active user may end up collecting a large number of media objects and corresponding tags. The meta-information stored in the tags – to which the user can add extra free-form information, mostly for personal use – will allow the user to manipulate the media collection in ways not commonly available before. With the improved media indexing advantages explained later, the user is strongly encouraged to use such tags.

When the user 'registers' their music, some of the tag information is embedded as meta data into the media file itself, however if a user has a media file with meta data and a tag, then where appropriate, the tag data locally overrides the embedded meta data. The degree to which tags information is embedded into the media is under user control - indeed, the value of the system to the user is that the user has full control over exposure of usage and other data.

Release of this information is encouraged, and rewarded through the improved searchability compared to current peer-to-peer systems, and through the use of 'freebies' and other marketing schemes. However, since no identifying information *ever* leaves the user, he or she is always free to 'turn off the tap' at any time with no penalty other than the loss of those extra benefits.

The value to the content provider/marketer/publisher is the possession statistics and the potential size of the population from which these are drawn. Their key value being not "Who has a particular tag?", but "Which sets of tags are commonly found together, and how can this information about the mix of music each user possesses be leveraged towards directed marketing?".

Importantly, this system also allows for the discovery and trial of new music and other media content without risk to the user. It primarily uses the traditional word-of-mouth recommendation from friends, but now users can trial the whole song at low-quality. This is in contrast to systems such as iTunes, where content must first be purchased, or as with the 30-second samples used by Amazon [3], (invariably the samples are unrepresentative of the song (it seems), as they were automatically produced). It is also possible to implement push-style media distribution without detracting from the rest of the system.

In contrast to some honour-based schemes such as Gnutella [8] and KaZaa [12], the 'free-riding' phenomenon [1] is avoided since all users are equal under this system in terms of the information transferred.

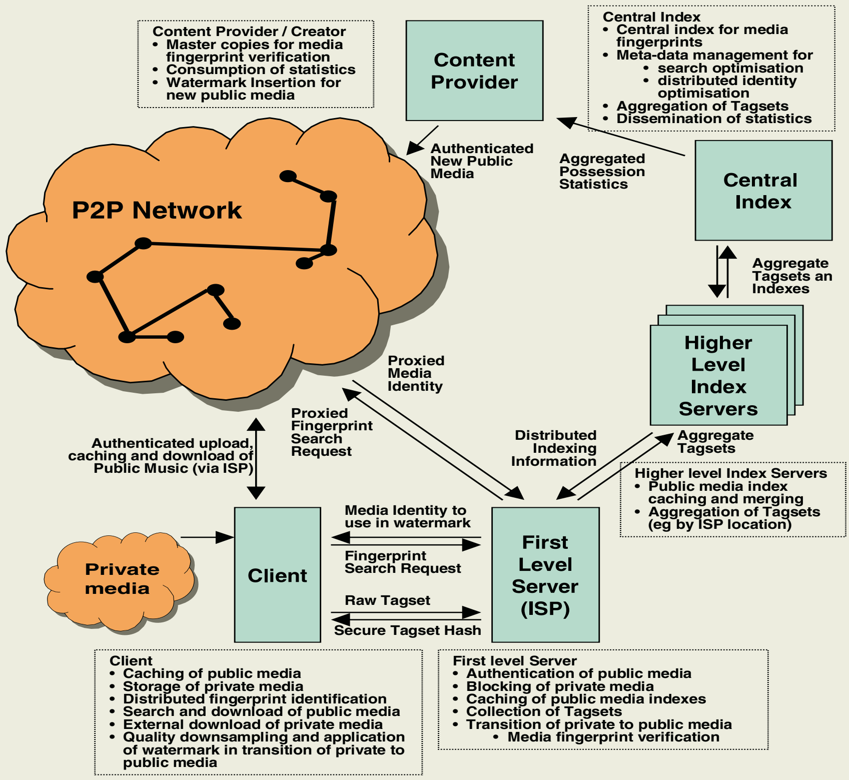

Content Provider / Creator
• Master copies for media fingerprint verification
• Consumption of statistics
• Watermark Insertion for new public media
Content Provider
Central Index
• Central index for media fingerprints
• Meta-data management for
• search optimisation
• distributed identity optimisation
• Aggregation of Tagsets
• Dissemination of statistics
Authenticated New Public Media
Aggregated Possession Statistics
Central Index
P2P Network
Aggregate Tagsets an Indexes
Higher Level Index Servers
Proxied Media Identity
Proxied Fingerprint Search Request
Authenticated upload, caching and download of Public Music (via ISP)
Distributed Indexing Information
Aggregate Tagsets
Higher level Index Servers
• Public media index caching and merging
• Aggregation of Tagsets (eg by ISP location)
Media Identity to use in watermark
Fingerprint Search Request
Raw Tagset
Secure Tagset Hash
Private media
Client
First Level Server (ISP)
Client
• Caching of public media
• Storage of private media
• Distributed fingerprint identification
• Search and download of public media
• External download of private media
• Quality downsampling and application of watermark in transition of private to public media
First level Server
• Authentication of public media
• Blocking of private media
• Caching of public media indexes
• Collection of Tagsets
• Transition of private to public media
• Media fingerprint verification

Fig. 1. Overview of the System

This system also fits with a full-purchase business model called Music2Share described by Kalker et al [13]. A user would use our system for free low and medium-quality media exchange in exchange for statistical information about which compatible media a user possesses and how that changes over time. This can be done without user identity being required.

In contrast, those users who want guaranteed high or maximum quality media may use the purchase model suggested by Music2Share.

2 Preliminary Design

The system is composed of a hierarchical set of media indexing servers, leading up to one or a small set of central servers which maintain a central media registry or index.

Figure 1 shows an overview of the system showing major protocol paths.

2.1 Media Types

There are three different types of media involved in this system. These correspond roughly to the similarly-named media types described by Kalker [13] in Music2Share (M2S).

Unregistered public media: This is public media which has not been registered by the client. Such media will play in a *very bandlimited* manner on compatible media players until registered, or on ordinary players which are not part of the system. Such media, while rendering badly, will still produce recognizable renditions. For example, music will render as if played through a telephone but it will still be recognisable and identifiable.

Registered public media: This is media that was obtained by the client from our system, or has been authenticated and registered by the client. It is unencrypted but bandlimited to a quality less than VHS or CD. A user who wants better quality must purchase the media (e.g. from M2S). Public media contains a high capacity watermark (such as in [19]), which is used for easy identification and contains the data required to render good quality audio/video. The watermark contains the lower and higher frequency components removed by the bandwidth limiting process of registration. A compatible player will decode the watermark and render it in good quality once the song is registered.

Encrypted public media: This is full bandwidth media of verifiable quality as per the M2S system. It is encrypted and the compatible media player can be instructed to silently pass the media to a M2S player when encountered. Note that this media type is not formally part of our system, but is compatible with it. It cannot be transferred as part of our system since M2S requires user identity.

Private media: This is media supplied by the user which has not been authenticated in any way. This media may come from anywhere, and the client is free to use it in any way. However private media will not be transmitted via our system and the client will not be able to take advantage of any indexing.

Authentication involves verifying the identity of the media from its 'fingerprint' (as with [13], then creating a down-sampled or bandlimited version of the media and inserting the watermark to create a public version of the media. This public version is then stored locally for P2P use.

Registration involves obtaining the identity and feature data of any new public media and combining it with the data collected from other public media in the client's possession (called a tagset), and then transferring the combined information to the media server. The server returns a secure hash of the tagset, and registration is complete.

Table 1. Media Types

Media Type	Quality	Usage	Bandwidth	Video size
Unregistered Public	Poor	Preview	6kHz	80x50
Registered Public	Public	Good Normal	15kHz	320x240
Private, Encrypted Public	Excellent	Premium	> 44kHz	full-screen

Thus users are free to distribute low ('telephone') quality music without registration, or good ('FM-radio') quality with registration. The registration process causes all submitted media files to be bandlimited to good quality, then watermarked / bandlimited to poor quality. In this way, high-quality music is never distributed, but the user is encouraged to use Music2Share to purchase high quality music if desired.

2.2 Media Index Servers and Possession Statistics

The Media index servers – typically operated by ISP's – are specially authenticated servers that form a federated information resource on tag movements, and this information, in the form of aggregated statistics files can be used by Copyright agencies to determine possession-based royalty payments.

The media publishers pay for the privilege of obtaining the statistical data for sets of songs from ISP's running an index server. The ISP can also derive income from advertising and various marketing-driven games like an online radio or TV station. Sets of tags (as playlists) can be delivered as part of a promotion.

It is expected that over time there would be a natural merging of server operations like the above over a few large ISP's, but ISP franchise arrangements could be used to distribute this load.

2.3 A Hierarchical Peer-to-Peer Approach

Many existing DRM-based music distribution systems such as iTunes [4] are based on a central server. This has the obvious disadvantage of high network load, poor scalability and net bandwidth bottlenecks, although this is not an insurmountable problem as shown by the architecture of Google [9]. One major difference is the average transaction size, which for Google is relatively small.

In terms of music sharing, the principal advantage of a centralised server model as in the original Napster [16] network model is the indexing. A central index can allow rapid match and complex search criteria as opposed to dynamically maintaining distributed indexes over many peers. By contrast, a P2P network has the advantages of being naturally scalable – distributing the network load and bandwidth, and providing a fault-tolerant manner of operation – any node may fail without major impact to the rest, as opposed to the single point of failure implicit in any client-server architecture. But a fully P2P system also has some distinct disadvantages: quality of service is uneven; slow propagation of information about new nodes; and sub-optimal searchability, especially with fuzzy search terms.

A compromise between a central server model and a fully decentralised model is therefore proposed: a hierarchical P2P (HP2P) network using distributed hash tables along the lines of Chord and TOPLUS [7].

A HP2P model combines the scalability advantages of P2P with the centralised indexing of a traditional client-server model. Indeed, a number of distributed indexing models exist, each with different search and network complexities (SearchTools [17], Collab [5]). Providing hierarchies allows for considerable search optimisation and may be sufficient to make fuzzy searches feasible.

Except for the largest ISP's, the first-level hierarchy would most likely encapsulate all the users for a particular ISP. In the second hierarchical level, multiple ISP-based song indexes would be combined. In other words, while the file-sharing would remain strictly P2P, the indexing would be more hierarchical.

If an ISP is distributed geographically, then there may be virtue in splitting on that basis. In this case, dynamic indexing may allow certain local meanings or interpretations to match more strongly, if the search request was made locally.

2.4 Users Sharing Music

Music is shared between users by exchanging the tags or the media (usually the former). The system will then automatically fetch either the corresponding tag or media, (at a network-friendly priority) whichever was not shared, in order to complete the pair, or derive a tag from the media if it was already suitably tagged with meta-data.

For Mobile-phone systems, using each mobile base transmitter as a media server can be a particularly efficient way to deliver media – bearing in mind the bandwidth / quality trade-off inherent in a mobile environment. Mobile users would then exchange tags very quickly and the media can arrive automatically at a network-friendly rate.

Tagsets can themselves be shared (in which case, they devolve to playlists) – allowing personalised music selection of new music – or created from a preliminary song-feature selection process. In this case, a new customised tagset can be regenerated periodically, and then distributed from a central repository, implementing a form of personally customised internet radio. This latter method nicely addresses one of the perennial questions in recommender systems based on usage – how to incorporate new music.

3 Conclusion

This short paper, briefly describes a HP2P system for sharing music where the user is able to share music freely, but where the marketing value of their ever-changing song collection is used to pay for it.

It also comments on how such as a system can interoperate with a fee-based P2P song-purchase system, to allow the user to easily purchase a higher quality rendition of the song initially obtained freely.

Acknowledgement

I would like to thank all the people in the User Needs Group of the Smart-Internet Cooperative Research Centre for their support.

References

1. Adar, E., Huberman, A., Free riding on gnutella. First Monday, **5** (2000) http://firstmonday.org/issues/issue5_10/adar/index.html. Accessed 9 May 2005.
2. Brilliant Digital Entertainment, Joltid, Altnet website http://www.altnet.com
3. Amazon Inc. any music search result on http://www.amazon.com
4. Apple Computer Inc, itunes Web Site, http://www.apple.com/itunes/
5. Collab.net Open Source Project, Project JXTA: A P2P Search tool, http://search.jxta.org/project/www/background.html, Accessed 9 May 2005.
6. MetaMachine Inc., eDonkey website, http://www.edonkey.com
7. Garces-Erice, L., Biersack, E., Felber, P., Ross, K., Urvoy-Keller, G., Hierarchical peer-to-peer systems. ACM/IFIP International Conference on Parallel and Distributed Computing (Euro-Par), Klagenfurt, Austria (2003) 1230–1239
8. Gnutella Open Source Project, Gnutella website http://www.gnutella.com,
9. Google Inc., Google website http://www.google.com/, Accessed 9 May 2005.
10. MarketWire Inc., Philips and Gracenote Launch Gracenote Mobile(SM) – First Global Music Recognition and Content Delivery Service for Mobile Phones, http://www.marketwire.com/mw/release_html_b1?release_id=61431, Article dated 9 January 2004
11. Hummel, T, Strømme Ø, La Salle, R (2003), Earning a Living among Peers the Quest for viable P2P Revenue Models, Proceedings of the 36th Hawaii International Conference on System Sciences, Hawaii (2003), p219
12. Sharman Networks, KaZaa website, http://www.kazaa.com
13. Kalker, T, Epema, D, Hartel, P, Lagendijk, R, van Steen, M, Music2Share: Copyright Compliant Music Sharing on P2P Systems, Proceedings of the IEEE, 92:6, (2004)
14. Microsoft Inc, MSN Music, http://music.msn.com
15. Musicmatch Inc, Musicmatch, http://www.musicmatch.com
16. Napster LLC., http://www.napster.com
17. Search Tools Consulting, Peer to Peer Searching, http://www.searchtools.com/info/peer-to-peer.html, Article dated 23 May 2002
18. Walmart Inc, Walmart Music Services, http://www.walmart.com/music_downloads/introToServices.do, Accessed 9 May 2005.
19. Xu, C., Wu, J.D. Feng, D., Content-Based Digital Watermarking for Compressed Audio, Proceedings of RIAO2000 Content-Based Multimedia Information Access, pp.390-402, Paris, France, (2000).

Analysis of Parity Assignment Steganography in Palette Images

Xinpeng Zhang and Shuozhong Wang

School of Communication and Information Engineering, Shanghai University
Shanghai 200072, P.R. China
{xzhang,shuowang}@staff.shu.edu.cn

Abstract. In parity assignment-based steganography for palette images, all colors in a host image are divided into two subsets, and each pixel is used to carry one secret bit. This paper describes an analytic method against the parity assignment-based steganographic techniques. By finding the rule of color modifications, a steganalyst can attempt to recover the original histogram in a way that is a reverse of data embedding. Because of the abnormal colors in the original image, an excessive operation will cause some negative values in the recovered histogram. This provides a clue for revealing the presence of secret message and estimating the length of embedded bit sequence.

1 Introduction

The objective of steganography is to send secret message under cover of a carrier signal [1]. It is generally accepted that any steganographic technique must possess two important properties: good imperceptibility and sufficient data capacity. The first property ensures that the embedded message is undetectable, and the second means efficiency in hidden communication. Despite that steganographic techniques only alter the most insignificant components of the host media, they inevitably leave detectable traces so that successful analysis, i.e., revelation of the presence of embedded data [2], is often possible. Many steganalytic techniques have been developed [3].

Various types of multimedia data can be used as carriers in steganography, among which palette images are popular since they are widely available and convenient to transmit via the Internet. Palette image uses a few, generally no more than 256, colors to provide acceptable visual quality. Each pixel possesses an index value mapped to a displayed color according to a palette, which includes all colors in the image. In the steganographic techniques for palette images proposed by Fridrich [4, 5], all colors in the palette are divided into two subsets representing respectively the secret bits 0 and 1. For a pixel into which one secret bit is embedded, no modification is needed if the original color belongs to the subset corresponding to the secret bit, otherwise the closest color in another subset is chosen to replace the original color. In [4], assignment of colors to the subsets is done according to the parity of the sum of red, green, and blue components.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1025–1031, 2005.

In [5], a smarter optimal parity assignment (OPA) method is used as described below:

1. Calculate the Euclidean distances between all pairs of colors $d_{ij} = |c_i - c_j|$, and arrange them in an ascending order to produce a sequence of distances, $\{d\}$. Set $C = \emptyset$. Iteratively repeat the next step until C contains all colors.
2. Orderly choose the distance d_{kl} from $\{d\}$ such that either $c_k \notin C$ or $c_l \notin C$. No such d_{kl} can be found if C already contains all colors. If neither c_k nor c_l belongs to C, pseudo-randomly assign c_k and c_l to the two different subsets according to a key. In case $c_k \notin C$ and $c_l \in C$, assign c_k into the subset that does not contain c_l. Update $C = C \cup \{c_k\} \cup \{c_l\}$.

In the OPA method, a color in the palette and its closest neighbor must belong to two different subsets [5]. Thus, the original color of any pixel is either kept unchanged or modified into its closest neighbor. The distortion introduced is therefore very small.

It is shown in this paper that both parity assignment steganographic techniques as mentioned in the above are not secure. A steganalyst can derive the rule of color modifications and try to recover the original histogram, and the negative value in recovered histogram provides a clue for revealing the presence of secret message and estimating the length of embedded bit sequence.

2 Steganalytic Method

Let us first study the effect of data embedding on a palette image. Denote colors in the image as c_1, c_2, ..., c_N, and divide them into two subsets using a parity assignment method. In OPA, a color and its closest neighbor must belong to two different subsets. But this cannot be guaranteed by using the method described in [4]. Figure 1 sketches a case of parity assignment when the palette contains 6 colors, in which the white and gray circles are used to represent colors in the two different subsets. In this figure, there are 6 arrows from c_j to c_i if c_i is the closest neighbor of c_j among all colors in the subset that does not include c_j. This means that the color c_j may be changed into c_i in data embedding. An $N \times N$ matrix, $\mathbf{A}$, is also used to indicate the rule of color modifications, in which an element $A(i, j)$ equals 1 if an arrow from c_j to c_i exists. Otherwise $A(i, j) = 0$. Clearly, there is only one element having a value 1 in each column, and the other $N-1$ elements are all 0. In the case of Figure 1,

$$\mathbf{A} = \begin{bmatrix} 0&0&0&0&0&0 \\ 1&0&1&0&0&0 \\ 0&1&0&1&0&0 \\ 0&0&0&0&1&0 \\ 0&0&0&0&0&1 \\ 0&0&0&0&0&0 \end{bmatrix} \tag{1}$$

Denote the numbers of color occurrences of the original image as h_1, h_2, ..., h_N, and those of the stego-image as h'_1, h'_2, ..., h'_N. Let α be the ratio between the number of embedded bits and the total pixel number of the cover image . Since the rate of color changes due to data embedding is approximately $\alpha/2$, h_ns and h'_ns are related by a matrix $\mathbf{T}(\alpha)$:

$$\begin{bmatrix} h'_1 \\ h'_2 \\ \vdots \\ h'_N \end{bmatrix} \approx \mathbf{T}(\alpha) \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_N \end{bmatrix} \tag{2}$$

where

$$\mathbf{T}(\alpha) = \mathbf{A} \cdot \alpha/2 + \mathbf{I} \cdot (1 - \alpha/2) \tag{3}$$

$\mathbf{I}$ is an identity matrix.

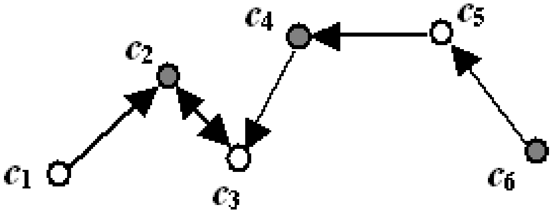


Fig. 1. A case of parity assignment when the palette contains 6 colors

If a steganalyst knows the parity assignment scheme used for data embedding, he can always derive the matrix $\mathbf{A}$. When the OPA method is used, although the steganalyst cannot find the particular division of two subsets due to the pseudo-random assignment in Step 2, he does know that a color may only be modified into its closest neighboring color in the embedding and the pseudo-random mechanism does not affect $\mathbf{A}$. Therefore the analyst can always work out the matrix. If the color assignment is determined by the sum of red, green, and blue components [4], it is even easier to obtain the matrix $\mathbf{A}$ since the steganalyst can find the exact subset division.

If a color c_s in the palette satisfies the following two conditions simultaneously: 1) occurrence of c_s in original image is very rare, i.e., $h_s \approx 0$, and 2) sum of occurrences of all c_r satisfying $A(s, r) = 1$ is significantly greater than h_s, we call c_s an *abnormal color*, which will provide a clue for detecting the presence of hidden data. As a simple example, when the original histogram is given in Table 1 and the color assignment in Figure 1, the color c_4 is an abnormal color. After data hiding with an embedding rate $\alpha = 0.5$, the histogram of the stego-image is also listed in Table 1. In fact, the procedure of data embedding is to change c_j to c_i in many pixels when $A(i, j) = 1$. Define

$$\mathbf{H}''(t) = \begin{bmatrix} h''_1(t) \\ h''_2(t) \\ \vdots \\ h''_N(t) \end{bmatrix} = [\mathbf{T}(t)]^{-1} \begin{bmatrix} h'_1 \\ h'_2 \\ \vdots \\ h'_N \end{bmatrix} \tag{4}$$

so,

$$\mathbf{H}''(\alpha) = \begin{bmatrix} h_1''(\alpha) \\ h_2''(\alpha) \\ \vdots \\ h_N''(\alpha) \end{bmatrix} \approx \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_N \end{bmatrix} \qquad (5)$$

Equation (5) indicates that $\mathbf{H}''(\alpha)$ is similar to the original histogram. It can be viewed as a procedure of changing c_i back to c_j when $A(i, j) = 1$. Thus, $h_s''(\alpha)$ approximately equals zero if c_s is an abnormal color. Considering $\mathbf{H}''(\alpha + \Delta t)$, where Δt is a small positive number, the value of $h_s''(\alpha+\Delta t)$ should be less than zero, since more pixels are subsequently departed from the abnormal color c_s. Figure 2 gives the value of $h_4''(t)$ as a function of t, which is derived from the stego-histogram in Table 1 and the matrix $\mathbf{A}$ in Equation (1). It is also shown that the curve of $h_4''(t)$ intersects the t-axis at $t = 0.53$, very close to the embedding rate 0.5.

Table 1. A sample of original and stego histograms

Colors c_i	c_1	c_2	c_3	c_4	c_5	c_6
Occurrence number h_i	52	186	467	9	742	144
Occurrence number h_i'	40	275	392	197	586	110

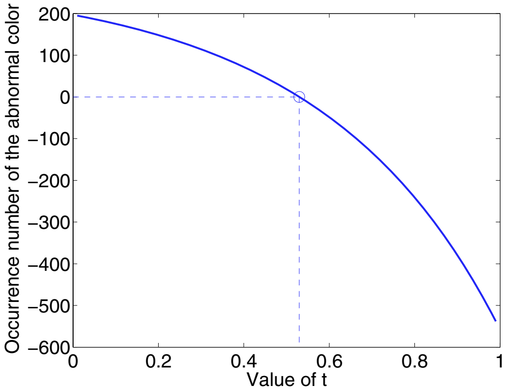


Fig. 2. Value of $h_4''(t)$ with different t

In natural palette images, there always exist some abnormal colors or semi-abnormal colors, where *semi-abnormal* means the two conditions in definition of *abnormal color* are roughly satisfied. Although a steganalyst does not know which color is abnormal or semi-abnormal since he does not have the original image, he can calculate the histogram of a suspicious image, obtain the matrix $\mathbf{A}$, and attempt to recover the original histogram in a manner that is the reversal of data embedding. Because of the presence of abnormal colors in the original image, an excessive operation will produce some negative values in the recovered histogram. This fact can be used for revealing the presence of secret message and estimating length of the embedded bit sequence. The detailed steganalytic method is as follows:

1. Get the histogram $[h'_1 \quad h'_2 \quad \ldots \quad h'_N]$ from a suspicious palette image.
2. Compute $\mathbf{H}''(t)$ with different t using Equation (4), and find the minimum value in $\mathbf{H}''(t)$,
$$y(t) = \min_{n=1,2,\ldots,N} \left[h''_n(t)\right] \tag{6}$$
3. Find the maximum value of t at which $y(t)$ changes from positive to negative as an estimate of embedding rate α_{E}. A small α_{E} implies a clear palette image, and a large α_{E} indicates the presence of hidden message.

While $y(\alpha)$ is always little greater than 0 as $\mathbf{H}''(\alpha)$ is similar to the original histogram, it is possible that $y(t) < 0$ when $t < \alpha$ for some special stego-images. So, in the steganalytic method, the last t at which $y(t)$ changes sign is taken as an estimate of embedding rate. Thus, from the stego-histogram in Table 1 and the matrix $\mathbf{A}$, one can obtain the estimated embedding rate = 0.53 (see Figure 3), which is close to the actual rate 0.50.

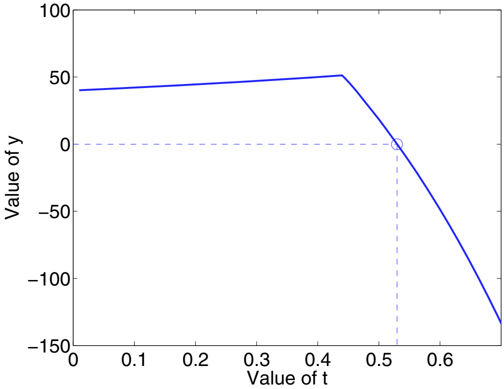


Fig. 3. Value of $y(t)$ with different t

3 Experiment and Discussion

Figure 4 is a stego-image with 256 colors and a size of 426×568, in which 80% pixels were used to carry secret bits embedded by OPA method, and its y-t curve is shown in Figure 5. The last intersection between the curve and the t-axis indicates $\alpha_{\mathrm{E}} = 0.82$, although one section of this curve on the left is below the t-axis.

In another experiment, a total of 80 palette images were used as covers, among which 40 were downloaded from the Internet and the others acquired using a digital camera in an uncompressed form and then converted to the gif

Fig. 4. Stego-image obtained by OPA steganography with $\alpha = 0.8$

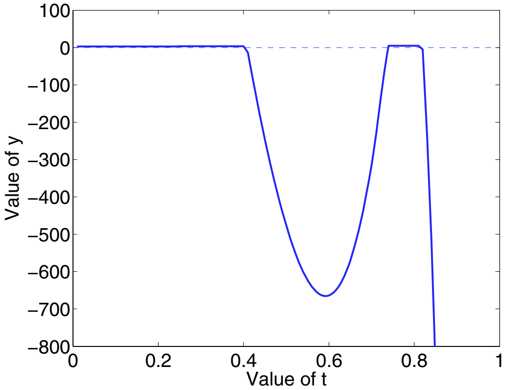


Fig. 5. y-t curve of the stego-image in Figure 4

format with a commercial tool. Both the OPA method [5] and the color assignment method based on the RGB sum [4] were performed to hide data with $\alpha = 0.5, 0.8$. Using the steganalytic technique described in the previous section, the estimated embedding rates of the original and stego images are illustrated in Figures 6 and 7. The two figures show that the more the embedding rate, the more accurate the estimation. Performance of analysis for the assignment steganography based on RGB sum is better than that for the OPA technique. For some images the estimates are considerably higher than the actual embedding rate. This is because that the colors in the covers do not exactly satisfy the abnormality conditions. In general, nonetheless, the proposed steganalytic method is effective.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 60372090) and Key Project of Shanghai Municipality for Basic Research (No. 04JC14037).

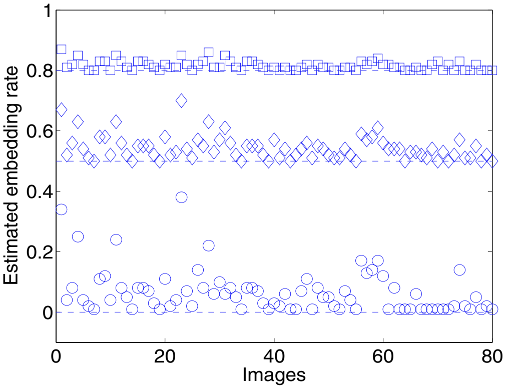


Fig. 6. Estimated embedding rates for OPA steganography (Circles: originals; Dimonds: stego-images with $\alpha = 0.5$; Squares: stego-images with $\alpha = 0.8$)

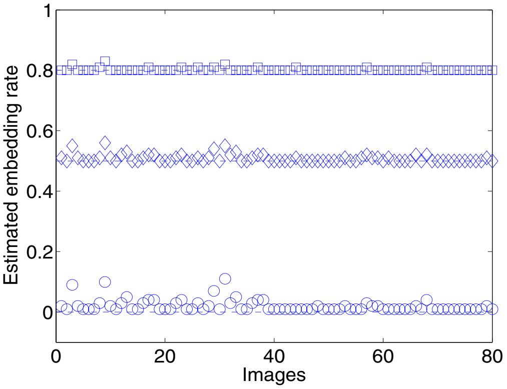


Fig. 7. Estimated embedding rates for the assignment steganography based on component sum (Circles: originals; Dimonds: stego-images with $\alpha = 0.5$; Squares: stego-images with $\alpha = 0.8$)

References

1. Petitcolas, F.A.P., Anderson, R.J., Kuhn M.G.: Information Hiding–A Survey. Proc. IEEE. **87** (1999) 1062–1078
2. Wang, H., Wang, S.: Cyber Warfare: Steganography vs. Steganalysis. Communication of ACM. **47**(10)(2004) 76–82
3. Fridrich, J., Goljan M.: Practical Steganalysis of Digital Images–State of the Art. Security and Watermarking of Multimedia Contents IV, Proceedings of SPIE. **4675**(2002) 1–13
4. Fridrich, J.: A New Steganographic Method for Palette-Based Images. Proc. of the IS&T PICS conference. (1998) 285–289
5. Fridrich, J., Du, R.: Secure Steganographic Methods for Palette Images. The 3rd Information Hiding Workshop, Lecture Notes in Computer Science. **1768** (2000) 47–60

A New Steganography Scheme in the Domain of Side-Match Vector Quantization

Chin-Shiuh Shieh[1], Chao-Chin Chang[1],
Shu-Chuan Chu[2], and Jui-Fang Chang[3]

[1] Department of Electronic Engineering, Kaohsiung University of Applied Sciences,
415 Chien-Kung Road, Kaohsiung, 807, Taiwan, R.O.C.
csshieh@cc.kuas.edu.tw

[2] Department of Information Management, Cheng Shiu University,
840 Cheng-Cing Rd., Niaosong Township, Kaohsiung County, 833, Taiwan, R.O.C
scchu@csu.edu.tw

[3] Department of International Trading, Kaohsiung University of Applied Sciences,
415 Chien-Kung Road, Kaohsiung, 807, Taiwan, R.O.C.
rose@cc.kuas.edu.tw

Abstract. This article reports a pioneer work on a steganography scheme in the domain of side-match vector quantization. The challenge associated with dynamic state code books had been resolved by two possible alternatives, namely code book partition by code words' mean and code book partition by pseudo random sequence. Experiment results reveal that imperceptibility required for secret communication can be ensured with the proposed approaches.

1 Introduction

Data secrecy had become an important issue as communication networks getting commoditization and widely spread, especially with the blooming of the Internet. Among others technologies, digital watermarking and steganography had received considerable attention in recent years for their theoretical and practical significance. Aimed at copyright protection, arbitration, and authentication, watermarking is the process of embedding extra information into a media clip. There have been a vast number of established methods [1]. However, it is still far from trivial to make the embedded watermark robust. Although closely related to digital watermarking, steganography has its own appeal for secret communication [2]-[3]. Steganography is the hiding of a secret message within an ordinary message and the extraction of it at its destination. Conventional cryptography techniques, such as DES and RSA [4], convert plain messages into random messages. Such a diffusion and confusion process tells potential attackers that there exist enciphered confidential message. Steganography takes cryptography a step farther by hiding an encrypted message in ordinal message so that no one suspects it exists. Ideally, anyone scanning your data will fail to know the existence of encrypted data [5].

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1032–1038, 2005.

The primary objective of steganography is the imperceptibility. That is, the carrier message should show no significant difference after the embedding of secret messages. In later sections, we will present a new steganography scheme in the side-match vector quantization domain. The proposed approach embeds a black-and-white icon into a gray-level carrier image, and maintains desired imperceptibility at the same time.

2 Side Match Vector Quantization

In order to reduce the bandwidth requirement for data transmission and space requirement for data storage, various data compression techniques had been developed [6]. Vector quantization (VQ) [7] is a widely adopted approach for lossy data compression. In applications regarding image, audio, and video, human sensory system is sophisticated enough to filter out limited data loose in the process of encoding and decoding. VQ and its descendents try to maintain high compression rate while retaining essential information carried in media clips.

The codebook size is a critical design parameter. It decides the encoding speed and the image quality. It seems that there is an inevitable tradeoff between these two important criteria. However, side-match vector quantization (SMVQ) [8] ingeniously tackles this dilemma by using dynamic, smaller code books for internal image blocks. A smaller code book implies fewer bits in encoding the indices. At the expanse of higher computational cost, overall compression rate can be significantly improved. This virtue makes SMVQ prevail in applications with limited bandwidth capacity.

With SMVQ, the image blocks in the first row and first column are dealt with in the same way as regular VQ. They differ in the processing of internal image blocks. Only a small portion of the main code book, called state code books, are used to encode internal image blocks. The state code books are selected from the main code book based on the surrounding information of the internal block to be encoded. Fig. 1 illustrates the general idea of SMVQ.

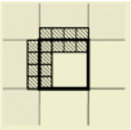


Fig. 1. Illustration for the processing of internal blocks with SMVQ

In Figure 1, blocks in gray denote blocks already encoded, and the block in bold border is the block under processing. 4-pixel by 4-pixel image blocks are used in the illustration. Let $X_{i,j}$ and $\hat{X}_{i,j}$ denote the original and encoded image blocks on the i-th row and the j-th column of the carrier image, and let $X_{i,j}(m,n)$ denote the gray value of the pixel on the m-row and n-th column within image block $X_{i,j}$. In encoding block $X_{i,j}$, we apply each code word, C_k,

$k = 1, \cdots, N$, in the main code book into the block in bold in Figure 1, and evaluate its associated side-match distortion (SMD) as follows:

$$SMD_{i,j,k} = \sum_{m=1}^{4} \left[\hat{X}_{i,j-1,k}(m,4) - C_k(m,1)\right]^2 + \sum_{n=1}^{4} \left[\hat{X}_{i-1,j,k}(4,n) - C_k(1,n)\right]^2 \tag{1}$$

N_s code words in the main code book with smallest side-match distortion are then picked up to form the state code book for the encoding of image block $X_{i,j}$. It is worthwhile to point out that the state code book is dynamic and state-dependent. It depends on the block to be encoded. Different internal image blocks have different state code books. More computational cost does incur in finding state-dependent state code books. However, the cost is justified if we consider the higher compression rate we can obtain.

3 Digital Watermarking in Vector Quantization Domain

Lu, et al. [9] had proposed a successful digital watermarking scheme which is robust to vector quantization compression. With their approach, the main code book is bi-partitioned into two sub code books of equal size using genetic algorithms. One of these sub code book is used to embed bit “0”, and the other is for bit “1”. That is, to embed a bit “1”, we find out the closest code word in sub code book “1”, and output its index. The partitioning serves to be a secret key. With the same information, the receiver end can then extract the embedded information. The primary objective of the partitioning is to minimize the extra error introduced in the embedding process.

For the purpose of information hiding, we will follow the same philosophy in dealing with those image blocks in first row or in first column. However, the same approach can not be applied to internal blocks directly, since the state code books for internal image blocks are state-dependent. It becomes computationally impractical in following Lu’s approach. In Section 4, taking both computational efficiency and image quality into account, we introduce two possible alternatives for the partitioning of the state code books.

4 Data Hiding in Side-Match Vector Quantization Domains

The functional block diagram for our approach to data hiding in SMVQ domains is given in Fig. 2. The embedding process for those carrier image blocks on first row or first column is different from that for remaining blocks. We will examine the detailed operation of each block in subsequent sub sections.

4.1 Pseudo-random Number Generators

For another level of secrecy, two pseudo number generators, namely Blum-Blum-Shub generator and linear congruent generator [10], are considered in our approach.

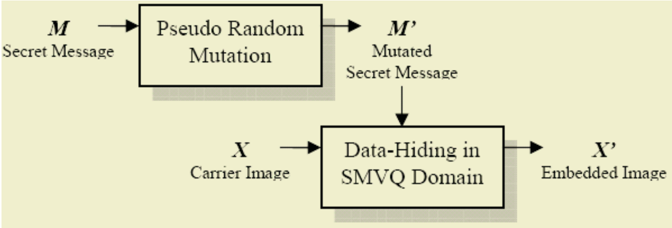


Fig. 2. Functional block diagram for the proposed approach

Based on squared residue, Blum-Blim-Shub generator (BBSG) can be used to generate pseudo random bit stream as follows:

$$\begin{cases} z_0 = k^2 \pmod{R} \\ z_r = z_{r-1}^2 \pmod{R}, r \geq 1 \end{cases}$$

and

$$b_r = z_r \pmod{2}, r \geq 0$$

where b_r, $r \geq 0$ is the desired pseudo random bit stream; k is a chosen number which has to be relative prime to R; R is the product of two prime number $R = p \cdot q$. It is required that $p \pmod{4} = q \pmod{4} = 3$. Taking the random seed k as secret key , BBSG is used in our approach to disturb the initial secret message.

Linear congruent generator (LCG) is capable of generating random number sequence within a specified range without repetition, as follows:

$$X_{n+1} = (P_1 X_n + P_2) \pmod{N}$$

where N is the specified range; P_1 and P_2 are two different prime numbers; X_0 is the random seed and serves as a secret key. LCG is used in our architecture for the partitioning of state code books as detailed in later sub-section.

4.2 Main Code Book Partitioning by Genetic Algorithms

For carrier image blocks on the first row or first column, we embed secret information into them with an approach similar to that of Lu, et al. [9]. We partition the main code book into two sub code books using genetic algorithms.

Genetic algorithms [11] solve optimization problems by encoding the problems into chromosome form. Chromosomes in initial population are randomly set. In each iteration, chromosomes are evaluated according to specified object function to determine their fitness values. Good chromosomes are selected and mated to generate new chromosomes by crossover and mutation. Worse chromosomes are replaced by the newly generated ones. As the evolution progressing, each chromosome competes with each other in order to survive. Good genes are supposed to survive and result in satisfactory solutions. Below is a pseudo code for genetic algorithms in general:

Pseudo Code for Genetic Algorithms

```
Randomize initialize population P(0)
WHILE terminate condition is not met
```

```
Evaluate P(t) using object function
Select P(t+1) from P(t) based on fitness value
Conduct crossover and mutation, on P(t+1)
```

The design of chromosome to match the problem at hand is the first decision to be made in applying genetic algorithms. In our system, each chromosome is a binary string of length N_M, which is the size of the main code book. It is required that there must be equal number of bit “0” and bit“1” in each chromosome string, so that it represents a legal main code book partitioning. In our problem, the fitness value is related to the extra error incurred by the embedding process. A chromosome is considered to have higher fitness value if the partitioning it represented introduces less encoding error.

4.3 State Code Books Partitioning

For image blocks other those on first row and first column, the encoding is based on their associated state code books. Due to the dynamic nature of state code books, pre-partitioning as we did for the main code book becomes infeasible. With a hop of maintaining desired image quality, we consider two possible solutions to this problem in this sub-section.

One possible solution is to partition the state code book according to the means of its code words. Once the state code book for a particular image block is constructed, we sort the code words in it according to their means, which is defined to be the average gray level of all its pixels. Then all code words in odd position are collected and form the sub state code book for the embedding of information bit “1”, and code words in even position are collected to form the sub state code book for the embedding of information bit “0”.

Another possible solution is to partition the state code book randomly. The linear congruent generator discussed in sub section 4.1 can be used to generate a random permutation of integers N_s, and therefore leads to a legal partitioning of the state code book. The random seed used in LCG now serves as a secret key and offer another level of secrecy.

Although both of above approaches do not take into account the extra encoding error incurred, satisfactory image quality can be expected if state code books of moderate size is used.

5 Experiment Results

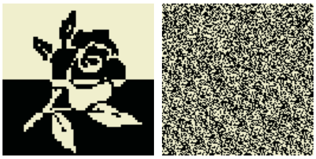
A series of experiments was conducted to verify the feasibility of the proposed approaches. Important parameters in the experiments are listed bellows:

- Carrier image: 512 by 512, 8-bit gray-level LENA
- Image block size: 4 by 4
- Secret message: 128 by 128, binary ROSE
- BBSG:$k = 151$,$R = 6691 \cdot 3467$
- Main code book size: 1025

- Main code book generation: LBG algorithm
- GA for main code book partitioning: Population size: 10;
  Crossover rate: 70 %; Mutation rate: 3 %
- LCG for state code book partitioning:$X_0 = 79$,$P_1 = 6691$,$P_2 = 3467$,$N = N_s$

For intended receiver, who has both the secret keys for BBSG and LCG, embedded secret message can be perfectly extracted, as shown in Fig. 3(a). Without knowledge on these secret keys, eavesdroppers can only obtain meaningless message as that in Fig. 3(b).

We also examined the effect of different state code book size. The results is summarized in Table 1. For state code book size of 256, there is only slight degradation in encoded image quality. The quality is even better for state code book size of 512. The embedded carrier images for state code book size of 256 are given in Fig. 4. It is difficult for human visual system to perceive the existence of secret message.



(a) (b)

Fig. 3. Extracted message (a) with secret keys, and (b) without secret keys

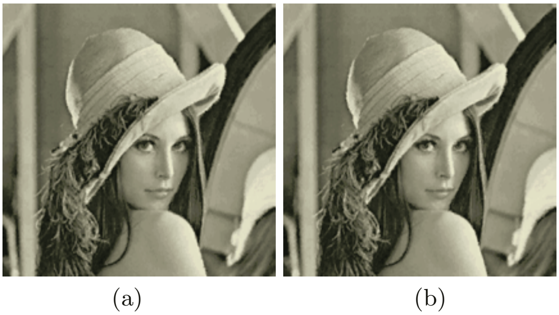
Table 1. Peak-Signal-Noise-Ratio (PSNR) value with respect to the original carrier image.

State Code Book Size	SMVQ	Data Hiding in SMVQ with State Code Book Partitioning by Mean	Data Hiding in SMVQ with State Code Book Partitioning by Random Sequence
256	30.339	29.91	29.75
512	30.983	30.53	30.62

6 Conclusions

A new steganography scheme in the domain of side-match vector quantization is proposed. The challenge associated with dynamic state code books had been resolved by two possible alternatives, namely code book partition by code words' mean and code book partition by pseudo random sequence. Experiment results reveal that imperceptibility required for secret communication can be ensured with the proposed approaches.

By its nature, the proposed scheme is robust to SMVQ compression. However, more investigation is required before we can have any conclusion on its robustness against general attachs.



(a) (b)

Fig. 4. Embedded carrier image with (a) state code book partitioning by code words' mean and (b) state code book partitioning by pseudo random sequence

References

1. Pan, J.-S., Huang, H.-C., and Jain, L. C., (Eds.), Intelligent Watermarking Techniques, World Scientific Publishing Company, Singapore, (2004).
2. Katzenbeisser, S. and Petitcolas, F. A.P., (Eds.), Information Hiding Techniques for Steganography and Digital Watermarking, Artech House Publishers, (2000).
3. Cole, E., Hiding in Plain Sight : Steganography and the Art of Covert Communication, Wiley, (2003).
4. Stallings, W., Cryptography and Network Security: Principles and Practice, 3rd Ed., Prentice Hall, (2002).
5. SearchSecurity.com at http://searchsecurity.techtarget.com
6. Sayood, K., Introduction to Data Compression, 2nd Ed., Morgan Kaufmann, (2000).
7. Gray, R. M., "Vector quantization", IEEE ASSP Magazine, pp. 4-29, (1984).
8. Chang, R.-F. and Chen, W.-T., "Image coding using variable-rate side-match finite-state vector quantization," IEEE Transactions on Image Processing, vol. 2, no. 1, pp. 104-108, 1993.
9. Lu, Z.M. and Sun, S.H., "Digital image watermarking technique based on vector quantization", IEE Electronics Online, vol.36, no.4, pp. 303-305, (2000).
10. BletchleyPark.net, Cryptograph, at http://www.bletchleypark.net/cryptology
11. Goldberg, D.E., Genetic Algorithms in Search Optimization and Machine Learning, Reading, MA: Addison Wesley, (1989).

Method of Hiding Information in Agglutinative Language Documents Using Adjustment to New Line Positions

Osamu Takizawa[1], Kyoko Makino[2], Tsutomu Matsumoto[3],
Hiroshi Nakagawa[4], and Ichiro Murase[2]

[1] National Institute of Information and Communications Technology
4-2-1, Nukuikita-machi, Koganei, Tokyo 184-8795, Japan
taki@nict.go.jp
[2] Mitsubishi Research Institute Inc.
3-6, Otemachi 2-chome, Chiyoda, Tokyo 100- 8141, Japan
[3] Graduate School of Environment and Information Sciences,
Yokohama National University
79-7, Tokiwadai, Hodogaya, Yokohama 240- 8501, Japan
[4] Information Technology Center,
University of Tokyo
7-3-1, Hongo, Bunkyo, Tokyo 113-0033, Japan

Abstract. Information hiding technology embeds information using the redundancy of information contained in cover data. Therefore, many information-hiding techniques for cover data with a lot of redundancy, such as images or sound signals, have been proposed. Most proposed information hiding techniques that set document to cover data tampered with layouts between spacing and words. In this paper, a new information hiding technique for agglutinative languages like Japanese or Korean that have no spaces between morphemes is proposed. By the proposed technique, digital documents are set to cover data and secret data is embedded by making the position of the new-line code inserted into document correspond to secret data. The technique can also be applied to plain text like an e-mail, which does not have layout information. Because the technique does not change the content of the cover data at all, the technique can be used not only as steganography aiming at performing secret communication but also as digital watermarking. Moreover, the technique has the feature whereby embedded data remains also in the printing output.

1 Introduction

Information hiding technology embeds information using the redundancy of information contained in cover data. Therefore, many information hiding techniques for cover data with a lot of redundancy, such as images or sound signals, have been proposed. On the other hand, with information hiding that sets document to cover data, secret data is embedded into cover data, i.e., cover text, and it is set to stego text. There is no redundancy in the character code of cover text, so it is difficult to embed secret data using character code. Therefore, with information hiding that sets document to cover data, there are many techniques of considering image documents, tampering with a layout, and embedding secret data [1]. In tampering with the layout, a

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1039–1048, 2005.

slight expansion or reduction of a line interval, a word interval, or character width, and slight rotation of a character are proposed. For example, if the number of standard pixels of a line interval is defined, the interval will be expanded in order to embed a bit "1", and the interval will be narrowed in order to embed a bit "0". Extraction of the embedded data will be performed by detecting expansion or reduction of the line interval by using a scanner. Therefore, the extraction success rate of embedded data will depend on the scanner resolution. If the grade of expansion or reduction is made small, an extraction error will increase instead of mental-fatigue-coming to be hard to tampering more.

On the other hand, the following researches exist for the technique of embedding secret data by character code into document instead of layout.

(1) SNOW
A technique of embedding secret data by setting English as the cover text and inserting space characters that are not visible on printed matter or a screen at the end of a line has been proposed. The name of the technique is SNOW [2]. SNOW is the technique of embedding 3 bits of secret data per line by inserting zero to seven space characters at the end of each line corresponding to secret data. By this technique, when the output system in which space characters are disregard or displayed as space is used, deterioration of the document does not occur. However, by using a certain kind of text editor, it can be seen that many unnatural space characters exist at the end of the lines. Moreover, a machine can discover the existence of unnatural space characters easily. When stego text is edited using an editor that erases excessive space characters, secret data is lost. Moreover, since the embedded information disappears, space characters with disregard or the output by output system displayed as space, SNOW is the limited technique which can be used only in the circulation as electronic data.

(2) Using the number of words of each line in a LaTeX source file
A technique that embeds secret data by adjusting the number of words of each line of English LaTeX source file by setting the source file to cover text has been proposed [3]. This technique uses the redundancy that the display document after compiling does not have information on the number of words of each line in the source file in general LaTeX. This technique is a kind of information hiding in computer program codes rather than information hiding in a document.

(3) FinPri.txt
A technique of embedding secret data, without changing the meaning of a text a lot is proposed by replacing words in the cover text with synonyms. The name of the technique is FinPri.txt [4]. This is the technique of using synonyms as redundancy of vocabularies. This technique, whose chief aim is saving rough meaning of the document, can be used for technical writing documents, such as a manual. Moreover, since an aggressor cannot discover the method of embedding secret data easily, there is the feature strong against the attack that removes secret data. However, there is a deterioration in documents in which importance is attached to delicate nuance when synonyms have been substituted, such as literary works or contracts. Moreover, since the alteration of vocabulary is an act that infringes on copyright, except when processed

by the author him/herself, there is a possibility of being restricted by law. Therefore, documents to which this technique can be applied are limited.

In addition, NICETEXT [5] and Texto [6] have been proposed. These methods are text generators that convert secret data into trivial or nonsense text. Therefore, they can't be applied to watermarking or fingerprinting for literary works.

Syntactically based and semantically based watermarking methods for natural language text have been proposed [7][8]. These methods require sophisticated natural language processing.

Linguistic steganography for Russian has been proposed [9]. The method is similar to the FinPri.txt mentioned above. In this method, words of the source text, i.e., cover text, are replaced with their synonyms. Therefore, as the text becomes deteriorated in comparison with the source text, the method is inapplicable to literary works. Moreover, the method needs a large synonymy dictionary and a huge collocation database.

The proposed technique in this paper does not insert invisible character code, and does not tamper with layout, but inserts new-line code that does not affect a document. In an agglutinative language like Japanese or Korean, it is comparatively free to start a new line in the middle of morphemes. That is, even if it starts a new line in the middle of morphemes, a hyphenation is not required. Then, secret data is embedded by making the position at which a new-line code is inserted correspond to secret data.

2 Outline of the Proposed Technique

The concept of procedure by the proposed technique is shown in Fig. 1. In the cover text the new-line code is contained only at the end of a paragraph, like that in a word processor document, and it is referred to as digital document by which the new-line code is not contained in the right end of every line on screen or paper. And the document which is inserted the new-line code for every line is set to stego text, making it correspond with the information to be embedded according to rules defined before-

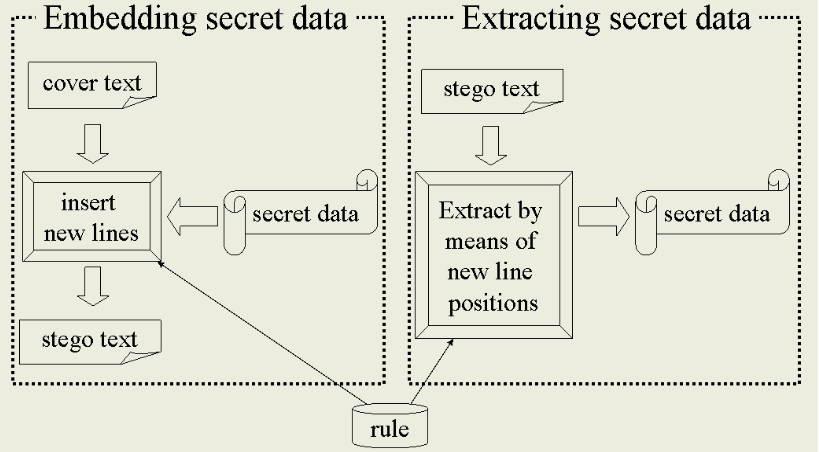
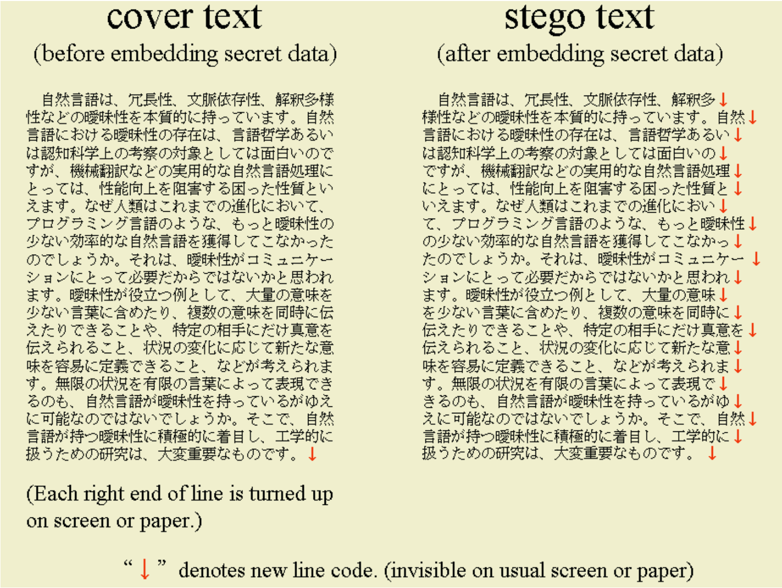


Fig. 1. Processing concept in which secret data is made to correspond to the position of a new-line code, embeds secret data into cover text, and is extracted from stego text

hand. In this technique the correspondence rule is a key for the encryption and decryption, and the same key is used for procedure both embedding and extracting secret data. This is a technique using being hard to distinguish the true new line with a new-line code from the turned up portion at right end of the line on general display system. An example of cover text and stego text is shown in Fig. 2.



cover text

(before embedding secret data)

　自然言語は、冗長性、文脈依存性、解釈多様
性などの曖昧性を本質的に持っています。自然
言語における曖昧性の存在は、言語哲学あるい
は認知科学上の考察の対象としては面白いので
すが、機械翻訳などの実用的な自然言語処理に
とっては、性能向上を阻害する困った性質とい
えます。なぜ人類はこれまでの進化において、
プログラミング言語のような、もっと曖昧性の
少ない効率的な自然言語を獲得してこなかった
のでしょうか。それは、曖昧性がコミュニケー
ションにとって必要だからではないかと思われ
ます。曖昧性が役立つ例として、大量の意味を
少ない言葉に含めたり、複数の意味を同時に伝
えたりできることや、特定の相手にだけ真意を
伝えられること、状況の変化に応じて新たな意
味を容易に定義できること、などが考えられま
す。無限の状況を有限の言葉によって表現でき
るのも、自然言語が曖昧性を持っているがゆえ
に可能なのではないでしょうか。そこで、自然
言語が持つ曖昧性に積極的に着目し、工学的に
扱うための研究は、大変重要なものです。↓

(Each right end of line is turned up on screen or paper.)

stego text

(after embedding secret data)

　自然言語は、冗長性、文脈依存性、解釈多↓
様性などの曖昧性を本質的に持っています。自然↓
言語における曖昧性の存在は、言語哲学あるい↓
は認知科学上の考察の対象としては面白いの↓
ですが、機械翻訳などの実用的な自然言語処理↓
にとっては、性能向上を阻害する困った性質と↓
いえます。なぜ人類はこれまでの進化におい↓
て、プログラミング言語のような、もっと曖昧性↓
の少ない効率的な自然言語を獲得してこなかっ↓
たのでしょうか。それは、曖昧性がコミュニケー ↓
ションにとって必要だからではないかと思われ↓
ます。曖昧性が役立つ例として、大量の意味↓
を少ない言葉に含めたり、複数の意味を同時に↓
伝えたりできることや、特定の相手にだけ真意を↓
伝えられること、状況の変化に応じて新たな意↓
味を容易に定義できること、などが考えられま↓
す。無限の状況を有限の言葉によって表現で↓
きるのも、自然言語が曖昧性を持っているがゆ↓
えに可能なのではないでしょうか。そこで、自然↓
言語が持つ曖昧性に積極的に着目し、工学的に↓
扱うための研究は、大変重要なものです。 ↓

" ↓ " denotes new line code. (invisible on usual screen or paper)

Fig. 2. Hiding information by inserting new-line codes in Japanese document. (Monospace fonts are used and each line is not justified in the stego text)

This technique has the following features.

(1) In this technique excessive codes like space characters are not inserted. The new-line code that is indispensable for documents is only used.
(2) If the output system does not disregard new-line code, the secret data remains also in printing output. This feature suggests this technique is applicable not only for electronic circulation but also for hard copy circulation.
(3) Since words are not replaced at all, the document does not deteriorate. Therefore, it can be applied to a literary work, or a contract, and it can be used as digital watermarking for it not only can be used as steganography, but also asserting the right of the work.

3 Detail of the Proposed Technique

3.1 Secret Data

Secret data is taken as a bit sequence of 0 or 1. To distinguish the line where secret data is embedded and not embedded at the stage of extraction, and to show the range of secret data, the flag sequences of the start and the end are used.

3.2 Proposal of Two Methods

The following two methods are proposed as correspondence rules.

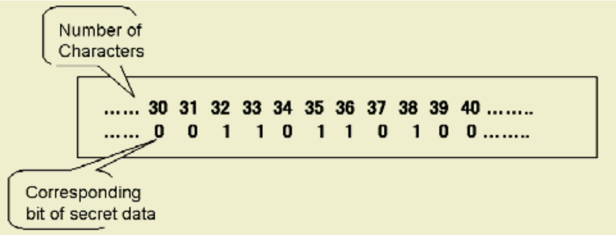
3.2.1 The Method to Make the Number of Characters per Line Correspond to Secret Data

The correspondence rule of this method defines the corresponding table of the number of characters per line, and the bit of secret data corresponding to it.

The embedding procedure is as follows. The new-line code is inserted in the position which becomes the number of characters per line corresponding to the bit of the secret data which is going to be embedded. However, priority is given to maintaining uniform line width in a general display system in case the number of characters per line is chosen from a corresponding table. Here, line width is defined as the total of the character width of all the characters in the line concerned. The Japanese character width of so-called "single-byte character" is defined as 1 per 1 character, and the character width of so-called "double-byte character" is defined as 2 per 1 character.

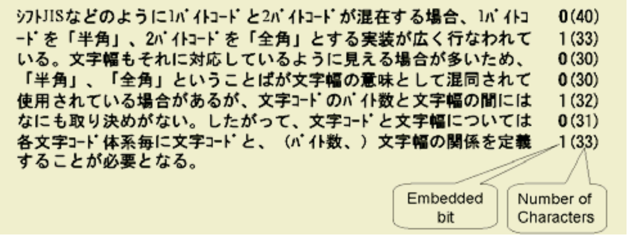
The extraction procedure is as follows. The number of characters of each line is counted and secret data is extracted using the same corresponding table.

An example of the corresponding table and an example of stego text generated by the table are shown in Figs. 3 and 4. In this system, single bit of secret data is embedded per line.



......	30	31	32	33	34	35	36	37	38	39	40	....
......	0	0	1	1	0	1	1	0	1	0	0	....

Fig. 3. Example of the rule table corresponding to number of characters of each line and single bit of secret data



ｼﾌﾄJISなどのように1ﾊﾞｲﾄｺｰﾄﾞと2ﾊﾞｲﾄｺｰﾄﾞが混在する場合、1ﾊﾞｲﾄｺ 0(40)
ｰﾄﾞを「半角」、2ﾊﾞｲﾄｺｰﾄﾞを「全角」とする実装が広く行なわれて 1(33)
いる。文字幅もそれに対応しているように見える場合が多いため、 0(30)
「半角」、「全角」ということばが文字幅の意味として混同されて 0(30)
使用されている場合があるが、文字ｺｰﾄﾞのﾊﾞｲﾄ数と文字幅の間には 1(32)
なにも取り決めがない。したがって、文字ｺｰﾄﾞと文字幅については 0(31)
各文字ｺｰﾄﾞ体系毎に文字ｺｰﾄﾞと、（ﾊﾞｲﾄ数、）文字幅の関係を定義 1(33)
することが必要となる。

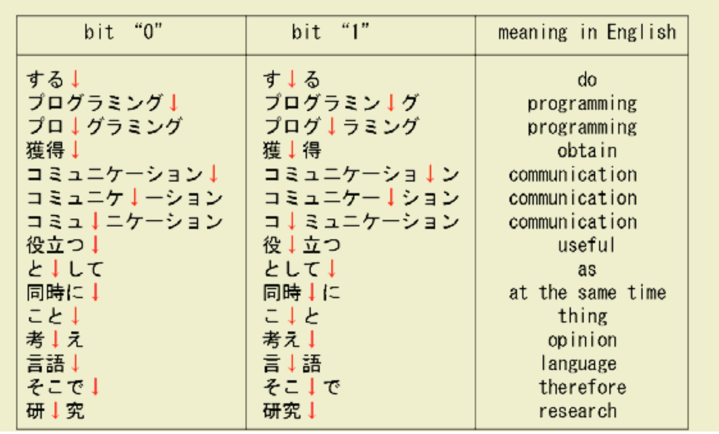
Embedded
bit
Number of
Characters

Fig. 4. Example of generated and justified stego text and embedded bits using the corresponding table shown in Fig. 3

The same corresponding table is used in procedure both embedding and extraction of secret data. Moreover, standard line width and minimum line width are defined, and they are used in procedure both embedding and extraction of secret data. Standard line width is a parameter for specifying standard line width to make the document of a favorite layout. Minimum line width is a parameter for not embedding bits of secret data in lines under the specified line width. It is made for line width not to embed secret data in a remarkable short line like the last line of a paragraph, or a caption by specifying minimum line width. Minimum line width serves as a required parameter in procedure of embedding and extraction.

3.2.2 The Method to Make New-Line Position in Morpheme Correspond to Secret Data

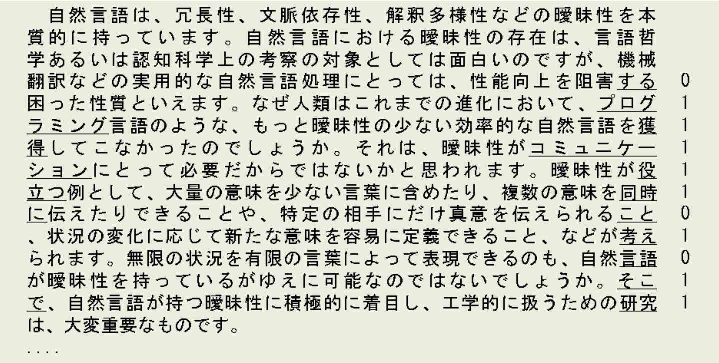
In this method the correspondence relationship between new-line position in each morpheme and the embedded bit is defined beforehand as the entry of a morphological-analysis dictionary, and secret data is embedded based on it. For example, if a word "suru" (do) is separated by a new-line code like "su|ru" ("|" denotes a new-line position), it will define corresponding to bit "1" of secret data. According to the standard line width specified at the stage of embedding procedure, secret data is embedded into morphemes that come near the end of the line. As shown in Fig. 5, long morphemes like "programming" and "communication" make two or more new-line positions correspond to bit "0" or bit "1", and it is made to start a new line with the number of characters which is not greatly different from the standard line width.



bit “0”	bit “1”	meaning in English
する↓	す↓る	do
プログラミング↓	プログラミン↓グ	programming
プロ↓グラミング	プログ↓ラミング	programming
獲得↓	獲↓得	obtain
コミュニケーション↓	コミュニケーショ↓ン	communication
コミュニケ↓ーション	コミュニケー↓ション	communication
コミュ↓ニケーション	コ↓ミュニケーション	communication
役立つ↓	役↓立つ	useful
と↓して	として↓	as
同時に↓	同時↓に	at the same time
こと↓	こ↓と	thing
考↓え	考え↓	opinion
言語↓	言↓語	language
そこで↓	そこ↓で	therefore
研↓究	研究↓	research

Fig. 5. Example of rule table corresponding to morphological headword and bit of secret data. (Each arrow denotes the place of new line code)

An example of stego text in which embedded information uses the corresponding table of Fig. 5 is shown in Fig. 6. Figure 6 is a justified document, and the variation in the number of characters of per one line can hardly be found visually. In the example of Fig. 6, the new-line positions of the first three lines are dummy in which secret data is not embedded, in the lines from the 4th to the 10th the flag sequence "0111110" is embedded, and at the 11th line the main part of embedded secret data "1011 ..." begins.



```
　自然言語は、冗長性、文脈依存性、解釈多様性などの曖昧性を本
質的に持っています。自然言語における曖昧性の存在は、言語哲
学あるいは認知科学上の考察の対象としては面白いのですが、機械
翻訳などの実用的な自然言語処理にとっては、性能向上を阻害する　0
困った性質といえます。なぜ人類はこれまでの進化において、プログ　1
ラミング言語のような、もっと曖昧性の少ない効率的な自然言語を獲　1
得してこなかったのでしょうか。それは、曖昧性がコミュニケー　1
ションにとって必要だからではないかと思われます。曖昧性が役　1
立つ例として、大量の意味を少ない言葉に含めたり、複数の意味を同時　1
に伝えたりできることや、特定の相手にだけ真意を伝えられること　0
、状況の変化に応じて新たな意味を容易に定義できること、などが考え　1
られます。無限の状況を有限の言葉によって表現できるのも、自然言語　0
が曖昧性を持っているがゆえに可能なのではないでしょうか。そこ　1
で、自然言語が持つ曖昧性に積極的に着目し、工学的に扱うための研究　1
は、大変重要なものです。
....
```

Fig. 6. Example of generated and justified stego text and embedded bits using the corresponding table shown in Fig. 5. (In order to show each morpheme shown in Fig. 5, the underlines are attached. The underlines are un-displaying in fact)

The method under this correspondence rule has the following features.

(1) Since how to Embedding per word can be defined, as compared with the method of Section 3.2.1, it is difficult to detect the rule of the correspondence relationship between the bit of embedding information, and a new line. Therefore, it is strong against an extraction attack.

(2) Since new-line position can be defined for every word, it is possible to carry out the definition that avoids a new line in unnatural position.

On the other hand, in order to achieve this technique, it is necessary to solve the problem of the extraction error caused by the error of morphological-analysis procedure, and the definition method in the case of a one-character morpheme.

4 Implemenation and Evaluation

The technique defined in Section 3.2.1 has been implemented on computer. The tool has been implemented using the Java language, which can run on various OS's. In this section the amount of embedded information is evaluated using the tool. In the proposed technique, bits of secret data are not embedded in the lines before the start flag, and the lines after the end flag. However, in this evaluation, the number of bits that can be embedded into all lines is defined as "the total amount of embedding".

In this technique, a corresponding table, standard line width, and minimum line width are specified as parameters on the occasion of embedding procedure. Moreover, on the occasion of extraction procedure, the same table and the same minimum line width as that used at the stage of embedding procedure are specified. In an agglutinative language, the position into which a new line is put is comparatively free. However, Japanese language has the procedure called "Japanese hyphenation" that avoids only a punctuation and a parenthesis being sent to the following line, and starting new line while being a number sequence is to avoid. If many restrictions about position of new line are defined, the document will be easier to read, namely, it will become a natural document. Instead, since the flexibility at the time of embedding secret data at the number of characters of each line narrows, the number of characters of each line varies greatly for every line, and the document will be an unnatural one. In order to compare this trade off, the following three methods have been implemented and evaluated according to the restrictions in the position of new line.

Method 1 – Priority is given to the homogeneity of line width.
In the method 1, except for restrictions in Japanese hyphenation, new lines are inserted so that each line of cover text may be in agreement with a standard line width as much as possible.

Method 2 – Inserting new lines in a specific type of character sequence are restricted.
The method 2 is a method using the restrictions that do not start a new line within a specific character sequence (numbers and alphabets) in addition to the restrictions in the method 1.

Method 3 – Priority is given to the boundary line of the same type of character sequence.
In Japanese there are three types of characters: Hiragana, Katakana, and Chinese characters. In addition to the restrictions in the method 2, the method 3 adds the restrictions that do not insert new lines into a character sequence of Chinese characters, Hiragana, or Katakana. Furthermore, if number of characters surrounded by parenthesises is 5 or less, new line code isn't inserted between the characters. Therefore, a great portion of new-line positions becomes the boundary line of character types (Chinese character/ hiragana/ katakana/ number/ alphabet). In Japanese, since the boundary of a character type e.g., between hiragana and a Chinese character, or between katakana and hiragana are the boundaries of clauses in many cases, and inserting new line in the boundary of clauses is desired from the viewpoint of the ease of reading.

In the corresponding table used for the evaluation, when line width is even, the bit of secret data is set to "1", and the bit is set to "0" when line width is odd. This is for measuring the amount of bits that can be embedded. In order to make an illegal decryption difficult, in an actual corresponding table the relationship between line width and a bit should be made random. The standard line width was set to 50 (equivalent to 25 double-byte characters), the minimum line width was set to 40, and the secret data was an 8-bit sequence "10110100." In this evaluation, the secret data was embedded repeatedly at all new lines excluding the line under the minimum line width from the beginning of a cover text.

The cover texts used for evaluation and each total rate of embedding, i.e., bandwidth, are shown in Table 1. Here, the total rate of embedding is defined as the value that divides the total number of embedded bits by 8 times of text size, which was converted into the value per bit.

According to the results of Table 1, the total number of embedding bits is proportional to the size of cover text mostly. Therefore, the total rate of embedding doesn't depend on the size of cover text, and it turns out that the total rate of embedding is almost fixed. Moreover, notably, the difference in the total rates of embedding by the types of text did not appear, and did not have most differences arising from the determination methods of a new-line position. Therefore, we can conclude to be the proposed technique whose amount of embeddable information is stable.

Table 1. Cover texts used for the evaluation and each total embedding rate

Type of Text		Size of Text (byte)	Genre or Title	the total number of embedded bits (unit is bit)			total rate of embedding(5)		
				Method 1	Method 2	Method 3	Method 1	Method 2	Method 3
News	General	1,929		36	37	34	0.23	0.24	0.22
		1,751		33	33	32	0.24	0.24	0.23
	Special field	2,258	News article about cryptology	39	40	38	0.22	0.22	0.21
		2,433	News article about MS-Windows	45	45	44	0.23	0.23	0.23
	For Kids	3,765	News commentaryfor Kids	68	68	58	0.23	0.23	0.19
Technical Paper	Special field	2,290	Technical Paper in Japanese Conference	40	40	39	0.22	0.22	0.21
		3,336	Technical Paper in Japanese Conference	62	66	62	0.23	0.25	0.23
Novel	Classical	3,789	“Makurano soshi”	74	80	70	0.24	0.26	0.23
		6,353	“Genji Monogatari”	121	123	114	0.24	0.24	0.22
	For Kids	3,606	“Alice's Adventure in Wonderland" (Japanese Edition)	72	73	59	0.25	0.25	0.20
		5,418	“Kaze no Matasaburo”	105	106	89	0.24	0.24	0.21
	General	5,640	“Wagahai wa neko dearu”	110	109	106	0.24	0.24	0.23
		1,866	“Rashoumon Gate”	36	39	34	0.24	0.26	0.23

5 Discussion and Conclusion

Differences in tampering with documents cannot be found between cover text and stego text when it is assumed that an attacker is able to get and compare both cover text and stego text, or to get and compare two or more stego texts. When it is assumed that attacker can get only a single stego text, the stego text needs to be as natural as possible. The proposed technique is effective in case that attacker can get only a single stego text.

In the proposed technique, if a correction that changes the position of new lines is made to a document, secret data will be eliminated. The threat of elimination is unavoidable, because to maintain naturalness so that the embedding will be hard to detect, the secret data has to be embedded only in new line codes. We consider that making it hard to detect the embedding is effective in reducing the threat of intentional elimination. The mechanical correction that arranges the number of characters of each line is made by mailer in many cases, the ends of the lines are only turned up and the original new-line positions are saved. In the proposed technique, the correction which deletes original new-line codes or regives new-line positions serves as a threat. Since the mechanical distinction with the text and titles of chapters or itemized statements is difficult, this correction is seldom made in application software for plain text. Therefore, a possibility that new-line positions may be changed by mechanical correction is considered to be infrequent. In general, attackers cannot tamper with hard copy. In the proposed technique, secret data remains also in hard copy. Moreover, if secret data is encrypted and embedded, the tolerance to extraction attack can be increased.

In the proposed technique, the message sender and recipient must share the same secret rule table. The characteristic is inconvenient especially for n-to-n communication. However in case that the technique is applied to fingerprinting, the problem doesn't occur, because recipient is the verifier, i.e., the sender.

As a future work, the total rate of embedding should be increased.

References

1. J.T.Brassil, S.Low, N.F.Maxemchuk, L.O'Gorman, “Electronic Marking and Identification Techniques to Discourage Document Copying,” Proc. IEEE INFOCOM '94, Vol.3, pp.1278-1287.
2. M.Kwan, “The SNOW Home Page,” See http://www.darkside.com.au/snow/.
3. T.Matsumoto, H.Itoyama, “Can Bypassing Lawful Access be Always Detected?,” Technical Report of IEICE, ISEC96-79, pp. 159-164, Mar. 1997 (in Japanese).
4. T.Matsumoto, H.Nakagawa, I.Murase, “Information hiding technical development for network-development of finger printing system for document -FinPri.txt,” Information-Technology Promotion Agency, Jun. 2000 (in Japanese).
5. M.Chapman, G.Davida, “Hiding the Hidden: A Software System for Concealing Ciphertext as Innocuous Text,” Proc. Int. Conf on Information and Communicatons Security, LNCS 1334, pp.335-345, Springer, 1997.
6. K.Maher, “Texto,” See http://www.eberl.net/textodemo.html.
7. M.J.Atallah, V.Raskin, M.Crogan, C.Hempelmann, F.Kerschbaum, D.Mohamed, S.Naik, “Natural Language Watermarking: Design, Analysis, and a Proof-of-Concept Implementation,” Proc Int. Workshop IH 2001, LNCS 2137, pp.185-199, Springer, 2001.
8. M.J.Atallah, V.Raskin, C.F.Hempelmann, M.Karahan, R.Sion, U.Topkara, K.E.Trizenberg, “Natural Language Watermarking and Tamperproofing,” Proc. Int. Workshop IH 2002, LNCS 2578, pp.196-212, Springer, 2002.
9. Igor A. Bolshakov, “A Method of Linguistic Steganography Based on Collocationally-Verified Synonymy,” Proc. Int. Workshop IH 2004, LNCS 3200, pp.180-191, Springer, 2004.

Hiding Biometric Data for Secure Transmission

Yongwha Chung[1], Daesung Moon[2], Kiyoung Moon[2], and Sungbum Pan[3]

[1] Department of Computer and Information Science, Korea University, Korea
ychungy@korea.ac.kr
[2] Biometrics Technology Research Team, ETRI, Daejeon, Korea
{daesung,kymoon}@etri.re.kr
[3] Division of Information and Control Measurement Engineering,
Chosun University, Korea
sbpan@chosun.ac.kr

Abstract. In this paper, we describe biometric watermarking techniques for secure user verification on the remote, multimodal biometric system employing both fingerprint and face information, and compare their effects on verification accuracy quantitatively. To hide biometric data with watermarking techniques, we first consider possible two scenarios. In the Scenario 1, we use a fingerprint image as a cover work and hide facial features into it. On the contrary, we hide fingerprint features into a facial image in the Scenario 2. Based on the experimental results, we confirm that the Scenario 2 is superior to the Scenario 1 in terms of the verification accuracy of the watermarked image.

Keywords: Biometrics, Biometric Watermarking, Multimodal Biometric System

1 Introduction

Traditionally, verified users have gained access to secure information systems, buildings, or equipment via multiple PINs, passwords, smart cards, and so on. However, these security methods have important weakness that can be lost, stolen, or forgotten. In recent years, there is an increasing trend of using **biometrics**, which refers the personal biological or behavioral characteristics used for verification[1].

In general, biometric verification/identification system can be divided into two modes: a **unimodal** biometric system which uses only a single biometric characteristic and a **multimodal** biometric system which uses multiple biometric characteristics. Building multimodal biometric systems has become an important research trend to overcome several limitations of unimodal biometric systems such as unacceptable performance in **large-scale** applications[1-5]. In this paper, both **fingerprint** and **face** are chosen as the biometric characteristics for our multimodal biometric system. In fact, there are many examples of multimodal biometric systems(such as border control[5]) employing both fingerprint and face information because fingerprint can provide a cost-effective, reasonable accurate solution and face can provide a non-intrusive solution.

Another issue in using biometrics in **remote applications** is how to protect the biometric information securely from unauthorized accesses. As mentioned above, biometric techniques have inherent advantages over traditional personal verification techniques. However, if biometric data(biometric raw data or biometric feature data) have been compromised, a critical problem about confidentiality and integrity of the

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1049–1057, 2005.

individual biometric data can be raised. For verification systems based on physical tokens, a compromised token can be easily canceled and the user can be assigned a new token. Similarly, user IDs and passwords can be changed as often as required. On the contrary, the biometric data cannot be changed easily since the user only has a limited number of biometric features such as one face and ten fingers.

To provide secrecy and privacy of the biometric data in remote applications, several techniques are possible such as cryptography and digital watermarking. The straightforward approach to guarantee the confidentiality/integrity of the biometric data is to employ the standard cryptographic techniques[6]. The biometric data is encrypted prior to be transmitted to the server, and the server can execute the verification procedure after decrypting the transmitted biometric data. The encrypted data is secured since it would be useless to a pirate without an appropriate key. Unfortunately, encryption does not provide secrecy once data is decrypted. In other words, cryptography can protect biometric data in transit, but once decrypted, biometric data has no further protection.

Digital watermarking can be considered as method for complement to cryptography. Digital watermarking places information, called *watermark*, within the content, called *cover work*. A watermark is never removed during normal usage and can be designed to survive various attacks such as decryption, re-encryption, compression, digital-to-analog conversion, and file format changes. Since watermarking involves embedding information into the host data itself, it can provide secrecy even after decryption. The watermark, which resides in the biometric data itself and is not related to encryption-decryption operations, provides another line of defense against illegal utilization of the biometric data[7]. For example, it can provide a tracking mechanism for identifying the origin of the biometric data. Also, searching for the correct decoded watermark information during authentication can render the modification of the data by a pirate useless, assuming that the watermark embedding-decoding system is secure. Furthermore, encryption can be applied to the watermarked data, combining the advantages of watermarking and encryption into a single system. As in [7], we will focus on watermarking only in the remaining of this paper.

In general, traditional digital watermarking techniques considered for many copy prevention and copyright protection applications may degrade the quality of the content, regardless of the difference between the original and watermarked versions of the cover work. Since watermarking techniques used to increase the security of the biometric data affects the verification accuracy of the biometric system due to that quality degradation, the effects on biometric verification accuracy need to be considered in applying the watermarking techniques.

In this paper, we consider possible two scenarios and evaluate the effects of each scenario on biometric verification accuracy. In the Scenario 1, we use the fingerprint image as the cover work and hide the facial information(e.g., eigen-face coefficients) into it. On the contrary, we hide the fingerprint information into the facial image in the Scenario 2. After implementing each scenario, we measure the effects of watermarking on biometric verification accuracy with various parameters.

The rest of the paper is structured as follows. Section 2 explains the overview of multimodal biometric systems, the attack points in remote applications, and previous biometric watermarking techniques. Section 3 describes the two scenarios considered to provide the secrecy in the multimodal biometric system, and the results of performance evaluation are described in Section 4. Finally, conclusions are given in Section 5.

2 Background

2.1 Multimodal Biometric Systems

A biometric verification/identification system can be divided into two modes: a unimodal biometric system which uses only a single biometric characteristic and a multimodal biometric system which uses multiple biometric characteristic. The multimodal system could be, for instance, a combination of fingerprint verification, face recognition, voice verification or any other combination of biometrics.

The multimodal biometrics has become an important research trend to overcome several limitations of the unimodal biometric system such as unacceptable performance and inability to operate on a large user population[8].

Multimodal biometric systems can be designed to operate in five integration scenarios : 1) multiple sensors, 2: multiple biometrics, 3) multiple units of the same biometric, 4) multiple snapshots of the same biometric, 5) multiple representations and matching algorithms for the same biometric[8].

Information presented by multiple traits may be consolidated at various levels. At the feature extraction level, the data obtained from each sensor is used to compute a feature vector. Integration at the feature extraction level is expected to perform better than fusion at two other levels. However, this is not always the best solution. The feature shapes of multiple biometrics may not be compatible and even if they are compatible there is still a problem of combining the feature set. Concatenation could result in a feature vector with a very large dimensionality. Fusion at the decision level is considered to be rigid due to the availability of limited information.

In this paper, we integrate two different biometric characteristics, face and fingerprint, to identify people.

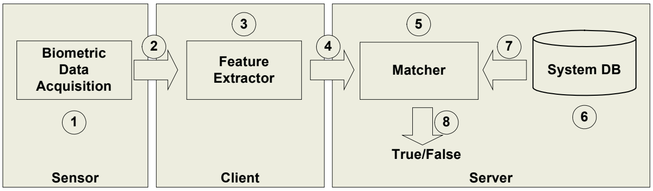


Fig. 1. Illustration of the Attack Points[6].

2.2 Attack Points

As shown in Fig. 1, many of the possible attacks in the biometric verification system were identified[2]: ① attack at the sensor, ② attack on the channel between the sensor and the feature extractor, ③ attack on the feature extractor, ④ attack on the channel between the feature extractor and the matcher, ⑤ attack on the matcher, ⑥ attack on the system database, ⑦ attack on the channel between the system database and the matcher, ⑧ attack on the channel between the matcher and the application requesting verification. Details of these attacks are explained in [2].

Note that the attacks ②,④,⑦ and ⑧ are launched against communication channels; they are also similar in nature and can be collectively called **"replay" attacks**[2]. In this paper, we focus on replay attack in the unsecure link of the multimodal biometric system.

2.3 Biometric Watermarking

Traditionally, Digital watermarking is a technique that hides a secret digital pattern, called a digital watermark, in a digital image, called cover work, or data.

Yeung and Pankanti[9] proposed a fragile invisible digital watermarking of fingerprint images based on a verification key that does not affect the recognition or retrieval accuracy in a fingerprint identification system. The fingerprints captured by the scanner are watermarked by the scanner and any tampering of the image data can be detected by the server using this method. Gunsel, Uludag and Tekalp[10] proposed a robust invisible watermarking of fingerprint images where the watermark can be verified even if the fingerprint image is cropped. Jain and Uludag[7] argued that when the feature extractor sends the fingerprint features to the matcher over an unsecure link, it may hide the fingerprint features in a cover work whose only purpose is to carry the fingerprint feature data.

As mentioned above, most previous researches about biometric watermarking use fingerprint image as cover work in order to increase integrity of the fingerprint image or hide a fingerprint feature data into unrelated cover work to guarantee the confidentiality of the fingerprint image.

In addition to capability of the watermarking technique, biometric watermarking must consider that verification performance based on watermarked images may not be inferior to the original non-watermarked images since embedding the watermark may change the inherent characteristics of the host image.

Ideal solutions of a digital watermark have been proposed in many papers. Each different watermarking technique has specific algorithm for the watermark embedding, but a general form of the watermark embedding algorithm is according to following equation

$$I_{WM}(x,y) = I(x,y) + k * W \,, \tag{1}$$

where $I_{WM}(x,y)$ and $I(x,y)$ are values of the watermarked and original pixels at location (x,y), respectively. The value of watermark bit is denoted as *W* and watermark embedding strength is denoted as *k*.

3 Hiding Biometric Data for Multimodal Biometric Systems

To hide biometric data using a fingerprint and a face, we first consider possible two scenarios. Then, we evaluate the performance of each scenario.

In the Scenario 1, we use a fingerprint image as the cover work and hide the facial information(e.g., eigen-face coefficients) into the fingerprint image. On the contrary, we hide the fingerprint information into the facial image in the Scenario 2. That is, in the equation (1), *W* is minutiae of the fingerprint image in the Scenario 1 and eigenface coefficients of the facial image in the Scenario 2.

3.1 Hiding Facial Features into Fingerprint Images

In the multimodal biometric system considered, two biometric data(fingerprint and facial images) are acquired from biometric sensors and the facial features(e.g., eigenface coefficients) are extracted to be used as watermarks. Finally, the embeddding site embeds the facial features into the fingerprint image. After detecting the facial features from the received watermarked fingerprint image, the detecting site calculates a similarity between the facial features stored and the facial features extracted. The detecting site also executes the fingerprint verification module with the watermarked fingerprint image received. The results of face and fingerprint verification modules can be consolidated by various fusion techniques such as majority voting to improve the verification accuracy of unimodal biometric systems.

Note that, in the Scenario 1, the accuracy of the face verification module is not affected by watermarking, whereas the accuracy of the fingerprint verification module can be affected by watermarking. To minimize the degradation of the accuracy of the fingerprint verification module, we use the embedding method proposed by Jain and Uludag[7]. That is, possible pixel locations for embedding are determined first by considering either minutiae or ridge information of the fingerprint image. Then, a bit stream obtained from the eigenface coefficients is embedded into the fingerprint image according to the possible pixel locations. This watermarking method, however, has several disadvantages. Because it is a type of *informed* watermarking, the minutiae or the ridge information is also needed in the detecting site.

We use our fingerprint verification algorithm[11] to evaluate the effects on fingerprint verification accuracy with the watermarked fingerprint image, and the result of evaluation will be described in section 4.

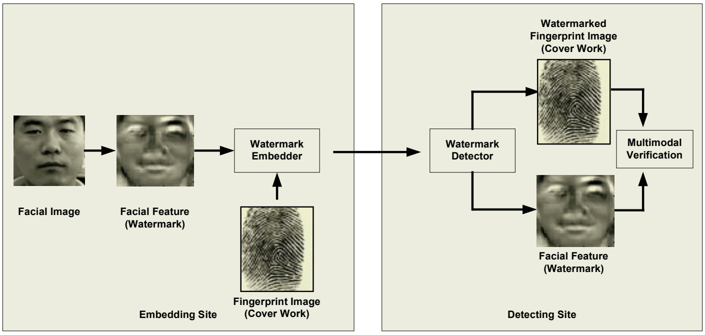

Facial Image
Facial Feature (Watermark)
Watermark Embedder
Fingerprint Image (Cover Work)
Embedding Site
Watermarked Fingerprint Image (Cover Work)
Watermark Detector
Multimodal Verification
Facial Feature (Watermark)
Detecting Site

Fig. 2. Diagram of Scenario 1

3.2 Hiding Fingerprint Features into Facial Images

In the Scenario 2, shown Fig. 3, the fingerprint features(e.g., minutiae) used as watermarks are extracted at the embedding site. Then, the minutiae are embedded into a facial image(cover work). After extracting the minutiae from the received watermarked image, the detecting site calculates a similarity between the minutiae stored and the minutiae extracted. The detecting site also executes the face verification module with the received facial image.

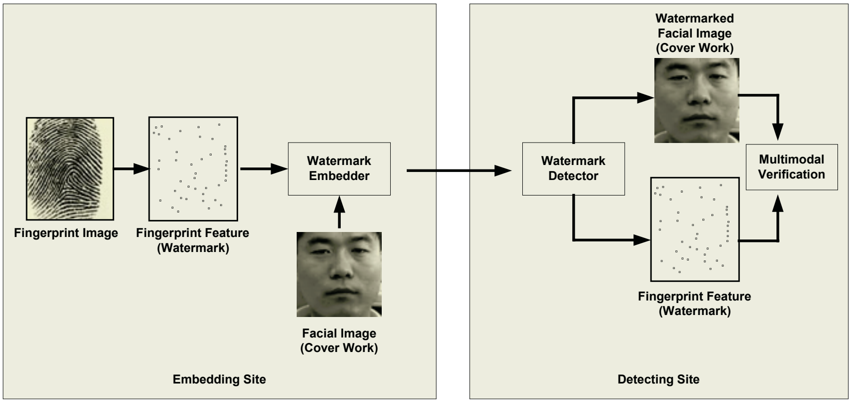

Fingerprint Image
Fingerprint Feature (Watermark)
Watermark Embedder
Facial Image (Cover Work)
Embedding Site
Watermarked Facial Image (Cover Work)
Watermark Detector
Multimodal Verification
Fingerprint Feature (Watermark)
Detecting Site

Fig. 3. Diagram of Scenario 2

To evaluate the effects on face verification accuracy with the watermarked facial image, two face verification algorithms are used: Principal Component Analysis(PCA)[12] and our face verification algorithm[13]. The PCA for face recognition is known as a global method[12] since it extracts facial features by using the bases describing a whole face. The bases are eigenvectors of the covariance matrix of the face images and can be regarded as face models, called *eigenfaces*. By projecting a face image onto the eigenfaces, the linear combination weights for eigenfaces are calculated. Then, these weights are used as a representation of the face. Although PCA method is simple and fast, there are some limitations in recognition under illumination and pose variations.

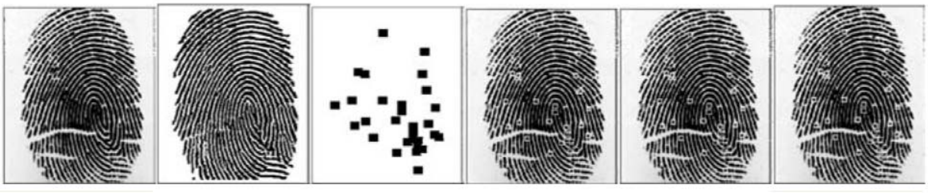
Our face verification method, called *Composite Template*, is based on Local Feature Analysis(LFA). LFA is known as a local method for face recognition since it constructs kernels detecting local structures of a face. However, LFA itself addresses image representation only and has some problems for recognition. By modifying LFA, we can get a new feature extraction method. Our method consists of three steps. After extracting local structures using LFA, we select a subset of them which is efficient for recognition. Then, the local structures are combined to represent them into a more compact form.

The result of evaluation that evaluates the effects on face verification accuracy with the watermarked facial image also will be described in the next section.

4 Performance Evaluation

To evaluate the effects on verification accuracy with watermarked images, we measure the accuracy of face and fingerprint verification with watermarked and non-watermarked biometric data. For the purpose of evaluating of the fingerprint verification accuracy, a data set of 4,272 fingerprint images composed of four fingerprint images per one finger was collected from 1,068 individuals by using the optical fingerprint sensor[14]. The resolution of the sensor was 500dpi, and the size of captured fingerprint images was 248×292.

As explained in section 3, we considered both minutiae and ridge information of the fingerprint image to decide the pixel locations for embedding. Fig. 4(a) represents an input fingerprint image and and 4(d) represents a fingerprint image with overlaid minutiae. Fig. 4(b) and 4(e) correspond to the ridge-based watermarking. That is, the input image shown in Fig. 4(a) is watermarked without changing the pixels shown in black in Fig. 4(b). Fig. 4(c) and 4(f) correspond to the minutiae-based watermarking. That is, the input image in Fig. 4(a) is watermarked without changing the pixels shown in black in Fig. 4(c). In the Scenario 1, the strength *k* in the equation (1) was selected as 0.06 because it is the most proper value for invisible watermarking as an experiment. Thus, human's eyes cannot distinguish any change of the fingerprint image caused by embedding watermarks.



(a) (b) (c) (d) (e) (f)

Fig. 4. Embedding Facial features to a fingerprint image: (a) an input fingerprint image, (b) fingerprint features based on the ridges, (c) fingerprint features based on the minutiae, (d) an input fingerprint image with overlaid minutiae, (e) a watermarked fingerprint image with overlaid minutiae generated from considering the ridge information of (b), (f) a watermarked fingerprint image with overlaid minutiae generated from considering the minutiae information of (c).

Fig. 5 shows four ROC(Receiver Operating Characteristic) curves for fingerprint verification. As shown Fig. 5, either the minutiae-based(represented as "Min Mask" in Fig. 5) or the ridge-based(represented as "Ridge Mask" in Fig. 5) watermarking methods introduce some degradation from the fingerprint verification accuracy of the non-watermarked case(represented as "Original" in Fig. 5) Note that both methods provided better accuracy than a method which did not consider the fingerprint features(represented as "No Mask" in Fig. 5).

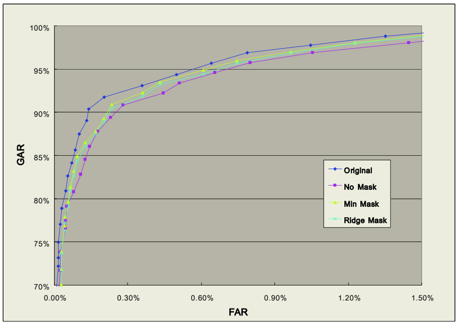
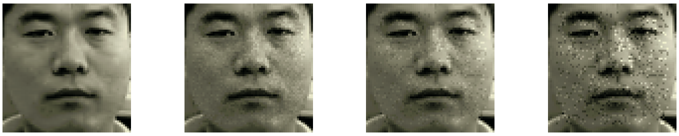
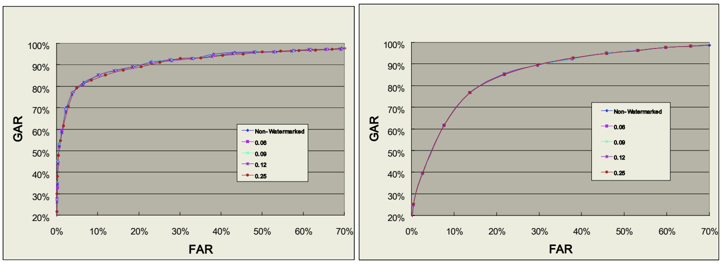


Fig. 5. ROC curves for the fingerprint verification performance



(a) input face image (b) *k*=0.06 (c) *k*=0.12 (d) *k*=0.24

Fig. 6. Facial images embedding the fingerprint features with various strength *k*



(a) (b)

Fig. 7. ROC curves for the face verification: (a) Composite Template using watermarked facial images with various strength *k*, (b) PCA using watermarked facial images with various strength *k*.

For the purpose of evaluating the face verification accuracy, a data set was collected from 55 people and composed of 20 facial images per one individual. The images from 20 people were used to construct bases for feature extraction, and the images from the rest were used for training(gallery) and test(probe). The size of images was 64×64, a simple Euclidean distance was adopted for computing a similarity.

To evaluate the effects on face verification accuracy with and without watermarking, we embedded the fingerprint feature into the facial image with various strength *k* explained in section 2. Fig. 6(a) represents an input facial image, and Fig. 6(b), 6(c) and 6(d) represent watermarked image with *k*=0.06, 0.12 and 0.24, respectively. With larger *k*(shown in Fig. 6(d)), a serious distortion of image quality can be found.

Fig. 7 shows ROC curves for face verification. In spite of serious distortion of image quality, the effects on verification accuracy with the PCA method are negligible. This is because the PCA method uses global information of face images. On the contrary, the Composite Template method using both local & global information of face images showed some degradation. However, the amount of degradation may be acceptable compared to the Scenario 1.

5 Conclusions

Biometrics are expected to be widely used in conjunction with other techniques such as the cryptography and digital watermarking on the network. For large-scale, remote user authentication services, the verification accuracy issue as well as the security/privacy issue should be managed. In this paper, the effects of watermarking on biometric verification accuracy were examined where both fingerprint and face characteristics were used simultaneously for accurate verification.

We first defined scenarios of the watermarking in the multimodal biometric system. Then, we analyzed the effects of watermarking in each scenario on the verification accuracy with various parameters. The experimental results show that the biometric features used as the watermark can be detected accurately in each scenario and the biometric verification accuracy is not affected by the detected biometric features. However, the verification accuracy of the biometric data used as the cover work may be degraded. Especially, the fingerprint verification accuracy in the Scenario 1 can be affected more sensitively than the face verification accuracy in the Scenario 2.

Acknowledgement

This research was supported by the MIC (Ministry of Information and Communication), Korea, under the Chung-Ang University HNRC-ITRC (Home Network Research Center) support program supervised by the IITA(Institute of Information Technology Assessment

References

1. A. Jain, R. Bole, and S. Panakanti, *Biometrics: Personal Identification in Networked Society*, Kluwer Academic Publishers, 1999.
2. D. Maltoni, et al., *Handbook of Fingerprint Recognition*, Springer, 2003.
3. R. Bolle, J. Connell, and N. Ratha, "Biometric Perils and Patches," *Pattern Recognition*, Vol. 35, pp. 2727-2738, 2002.
4. B. Schneier, "The Uses and Abuses of Biometrics," *Communications of the ACM*, Vol. 42, No, 8, pp. 136, 1999.
5. P. Verga, "DoD Biometrics and Homeland Defense," *Proc. of Biometric Consortium Conference*, 2004
6. W. Stallings, Cryptography and Network Security, Pearson Ed. Inc., 2003.
7. Jain A.K., Uludag U., and Hsu R.L., "Hiding a Face in a fingerprint Image," *Proc. of Int. Conf. On Pattern Recognition*, vol. 3, pp. 756-759, 2002.
8. A. Ross and A. K. Jain, "Multimodal Biometrics: An Overview," *Proc. of 12th European Signal Processing Conference (EUSIPCO)*, pp. 1221-1224, 2004.
9. Yeung M. and Pankanti S., "Verification Watermarks on fingerprint Recognition and Retrieval," *Journal of Electronic Imaging*, vol. 9, no. 4, pp. 468-476, 2000.
10. Gunsel B., Uludag B., and Tekalp A.M., "Robust Watermarking of Fingerprint Image," *Pattern Recognition*, vol. 35, no. 12, pp. 2739-2748, 2002.
11. S. Pan, et al., "A Memory-Efficient Fingerprint Verification Algorithm using A Multi-Resolution Accumulator Array for Match-on-Card," *ETRI Journal*, Vol. 25, No. 3, pp. 179-186, 2003.
12. Turk, M.A., Petland, A.P., "Face recognition using eigenface," *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Maui, Hawaii (1991)
13. Y. Lee, et al., "Local and Global Feature Extraction for Face Recognition," *To be published on Proc. of Int'l Conf on Audio- and Video-Based Person Authentication 2005(AVBPA), New York, USA, July 20-22, 2005*
14. NiGen, http://www.nitgen.com

VQ Image Steganographic Method with High Embedding Capacity Using Multi-way Search Approach

Chin-Chen Chang[1, 2], Chih-Yang Lin[2], and Yu-Zheng Wang[2]

[1] Department of Information Engineering and Computer Science,
Feng Chia University, Taichung, Taiwan, 40724, R.O.C.
ccc@cs.ccu.edu.tw

[2] Department of Computer Science and Information Engineering
National Chung Cheng University, Chiayi, Taiwan, 621, R.O.C.
{gary,wyc92}@cs.ccu.edu.tw

Abstract. Embedding large quantities of data in VQ (Vector Quantization) images is a thorny problem, since the hiding schemes usually have to change the index values of the VQ images, which might cause serious image distortion. As a result, many currently existing methods can only afford to support a small embedding capacity. In this article, we shall propose a new method that uses the genetic clustering technique on the codebook to obtain better clusters so that the replacement distortion of indices can be reduced. Then, we apply multi-way search to hide the secret data. Experimental results show that our new method outperforms existing schemes on both image quality and embedding capacity.

Keywords: Steganography, data hiding, clustering, multi-way search

1 Introduction

Image steganography is the art of conveying messages in a secret way that only the receiver can decrypt the message. Since steganography is to avoid secret data from being revealed from the stego-image without proper authorization, two main issues should be addressed: payload and resistance. However, these two requirements come in a tradeoff situation. Conventionally, the formal problem is referred to as *data hiding* and the later is known as *watermarking* [1, 9]. In this paper, we only focus on the former problem.

To do steganography, the simplest way is probably the least-significant bit (LSB) insertion method [3, 15], which embeds secret data in the least-significant bits of the stego-image and usually employs a pseudo-random number to clutter the embedding order to achieve security [5, 9]. This method changes the pixel values slightly, which would not result in a big difference in the intensity.

Although the LSB method is simple, it is not suitable for VQ-based steganography. VQ (Vector Quantization) is a lossy compression method [7, 8] that uses a codebook with codewords in it to represent an image. In recent years, many methods have been proposed to hide secret data in the VQ system. In 1999, Lin et al. proposed an LSB-like method for VQ [11], where one secret bit is embedded in one image block represented by a codeword index. They partition a codebook into two equally-sized sub-codebooks. The sub-codebooks are rearranged by the Pairwise Nearest Clustering

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1058–1064, 2005.

Embedding (*PNCE*) method [11] such that all pairs of codewords between the sub-codebooks are as similar as possible and are located at the same positions. After that, one sub-codebook is responsible for the odd indices and the other is responsible for the even indices. This way, Lin et al's method can embed the secret data in the least significant bits of the codeword indices.

In 2003, Du and Hsu [6] proposed a high-capacity embedding method based on VQ without applying LSB-like approaches. The method uses hierarchical clustering to partition the codebook into several clusters and translates the secret data into a decimal number which is represented by the combination of elements in the clusters that the block indices belong to. However, the image quality produced by this method is not desirable because it heavily depends on the clustering result. A larger embedding capacity comes from larger clusters, which will significantly degrade the image quality.

In this paper, we shall improve Du et al's method to make it more effective in embedding greater quantities of data in VQ images while keeping a high image quality. Instead of using the hierarchical clustering technique involved in Du et al's method, we apply genetic clustering for better clustering results. In addition, we shall also propose two embedding schemes, binary search and multi-way search, to conceal the secret data. The multi-way search is the generalized version of the binary version, and these search strategies are efficient to find the secret data when the secret data falls near the search boundaries. The experimental results show that our methods do better than others in terms of image quality and embedding capacity.

2 The Proposed Hiding Scheme on VQ

The proposed method firstly clusters the codebook by using a genetic algorithm. Then the search scheme is applied.

2.1 Preprocessing of Codebook

The first step we take here is clustering the codewords from the codebook. The purpose of this preprocessing procedure is to find a similar codeword efficiently in the same cluster for later usage in the searching procedure. The clustering result has a strong impact on the stego-image quality and the embedding capacity. Therefore, we apply the *GA-clustering* method [14] based on *k*-means using the genetic algorithm for obtaining optimal clustering results. This method has three main advantages. First, *k*-means discovers clusters in spherical shapes with better cohesion within a cluster than clusters with arbitrary shapes. Second, the genetic algorithm helps the *k*-means method toward optimal solution with the number of clusters to be grouped. Third, the clustering results by the *GA-clustering* algorithm are irrespective of the starting configuration, which is the choice of the initial cluster center.

The basic operations genetic algorithms execute are selection, crossover, and mutation, which are shown in Fig. 1. In the following, let's have a brief look at the *GA-clustering* technique.

To perform genetic operations, the chromosome and the fitness function should be defined first [13]. In *GA-clustering*, each chromosome is composed of a sequence of

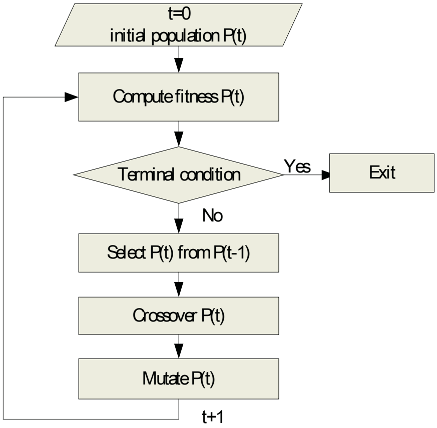


Fig. 1. Basic operations of genetic algorithms

real numbers representing the centers of the clusters. Assume there are g clusters C_1, C_2,…, C_g with centers z_1, z_2, …, z_g selected from n input points x_1, x_2, …, x_n. The fitness function f of *GA-clustering* is defined in Eq. (1), and the chromosome with the maximum fitness value is the best clustering result when the genetic operations are finished.

$$f = 1/M, \text{where } M = \sum_{i=1}^{g} M_i \text{ and } M_i = \sum_{x_j \in C_i} \left\| x_j - z_i \right\| \tag{1}$$

GA-clustering starts with p chromosomes as the initial population. The length of each chromosome is fixed during the genetic operations. The selection operation selects p chromosomes from the mating pool by using the roulette-wheel method [13], where the higher the fitness value is, the larger the probability of the chromosome to be selected. The crossover operation exchanges information between two parent chromosomes at crossover rate μ_c to generate two descent chromosomes, or called offspring. In addition, the mutation operation changes a cell's value of a chromosome at mutation rate μ_m within a fixed range defined by Eq. (2).

$$\begin{cases} v \pm 2 * \delta * v, \ v \neq 0. \\ v \pm 2 * \delta, \ v = 0 \end{cases}, \tag{2}$$

where v is a center value of a chromosome.

The three genetic operations are executed iteratively until the maximum fitness value hardly changes. After termination, the codewords of the codebook are clustered according to the chromosome with the maximum fitness value.

2.2 Binary and Multi-way Search Schemes for Embedding the Secret Data

After clustering, assume the codebook is partitioned into g clusters C_1, C_2,…, C_g and $C_i \cap C_j = \phi, \forall i \neq j$. The codewords in each cluster are sorted in non-decreasing order by PCA (principal component analysis) projection [4, 10]. The power of PCA is projecting a higher-dimension input vector onto a lower-dimension space while still preserving the maximal variances of the input vectors on the new coordinate axes. Each codeword in our scheme is projected onto a one-dimensional space in order for sorting to be conveniently done. According to the definition of VQ, the VQ image is represented by the indices of the codebook. We denote each index value of the VQ image as cw_{ij}, which means the codeword is in the j-th index of the i-th cluster.

Since the secret data S represented by a bit stream can be regarded as an unsigned integer I in $[2^k, 2^{k+1}]$, where $k = \lfloor \log_2 I \rfloor$. In the beginning, the value of k is embedded into the head of the stego-image by Du et al's method [6] to denote the lower bound of I. After that, the proposed binary search scheme is applied to search the remaining value of $(I - 2^k)$ in the range of $[0, 2^k]$. To distinguish between the value of k and the value of $(I - 2^k)$ embedded in the stego-image, we set a break point bp_i between them. The bp_i can be the median or mean codeword of the i-th cluster. In the proposed method, we set bp_i as the mean codeword with the value of $(\frac{\sum_{j=1}^{n} x_{ij}^1}{n_i}, \frac{\sum_{j=1}^{n} x_{ij}^2}{n_i}, \ldots, \frac{\sum_{j=1}^{n} x_{ij}^d}{n_i})$, where n_i is the number of codewords in the i-th cluster, and x_{ij}^d is the d-th element of the j-th codeword in the i-th cluster. Here, each bp_i should be unique in the codebook. Set the break point to be the mean codeword of the corresponding cluster can improve the embedding capacity but only slightly affects the compression rate.

The binary search directions are indicated by cw_{ij}: if j is odd, the search direction is to the right; otherwise, it is to the left. For example, assume the value of $(I - 2^k)$ is v'. If v' is closer to 0, then we reset the search range for v' to $[0, 2^{k-1}]$ and change the corresponding j of cw_{ij} to be even. By contrast, the range is to be changed from $[0, 2^k]$ to $[2^{k-1}, 2^k]$, and the corresponding j of cw_{ij} gets reset to be odd if v' is closer to 2^k. This completes one iteration session of the binary search. Note that (1) the cw_{ij} is ignored if the size of cluster i is one, since it does not help with the search; (2) the modification of j is done by decreasing or increasing it just by one, since either cw_{ij-1} or cw_{ij+1} would be the codeword closest to cw_{ij} in the sorted cluster.

The above process is repeatedly applied to the new range to search v' until the following objective function f is minimized:

$$f(n, l', u') = n + \log_2\left(\min\{(v'-l'), (u'-v')\}\right), \tag{3}$$

where n is the number of search operations performed, and ℓ', u' are the lower bound and upper bound of the new ranges, respectively. When the binary search is done, a break point is laid, and then the value of v', within $[\ell', v'-\ell']$ or $[u'-v', u']$, can be pointed out by Du et al's method. Finally, two break points are set to denote the end of the hiding process to get ready for the next piece of data to be hidden.

The binary search scheme can be generalized as a multi-way search scheme. Here, we provide two versions of multi-way search: fixed multi-way search and adaptive multi-way search. Fixed multi-way search uses fixed search branching factors; therefore, the index cw_{ij}^x of the cover image cannot be the embedded data if the size of cluster i is smaller than the fixed branching factors. In such a case, we unite cw_{ij}^{x}'s following indices such that the product of these corresponding clusters' sizes is the smallest number that is greater than or equal to the fixed branching factors. On the contrary, the number of branching factors in adaptive multi-way search depends on the cluster size that cw_{ij}^{x} belongs to.

3 Experiments

The experiments use three standard 512 × 512 gray level images, "Lena", "F16", and "Pepper" shown in Fig. 2, as the cover images to hide a random bit-stream in. The codebook of size 512 used in the experiments was generated by the LBG algorithm [12].

(a) Lena (b) F16 (c) Pepper

Fig. 2. Three cover images

Table 1 shows how the PSNR value compared between our methods and Du et al.'s method after the embedding of the fixed size random data (kbits). The results indicate that the proposed methods always do better than Du et al.'s method when the same random data size is embedded. The PNNE (pairwise nearest-neighbor embedding) method [11], on the other hand, applies the nearest-neighbor rule to pair the closest codewords, resulting in a small embedding capacity. In addition, Table 2 shows the results the fixed multi-way search method provided given the same cover images and secret data. Fig. 3 shows the relationship between the number of clusters and the corresponding PSNR values when the secret random data size varies (the values above the curve).

Table 1. Performance comparison of PSNR values of various methods

Embedding bits	PNNE		Du et al.'s method		Binary search		adaptive multi-way search	
	Lena	F16	Lena	F16	Lena	F16	Lena	F16
8K	29.61	29.61	30.74	30.09	31.98	31.45	31.99	31.45
16K	28.35	27.98	29.04	28.42	31.40	31.12	31.42	31.24
32K	NA	NA	28.58	28.08	30.27	30.16	30.29	30.16
48K	NA	NA	27.61	27.40	28.63	28.20	28.68	28.19
64K	NA	NA	25.99	26.14	26.53	26.34	26.58	26.36
80K	NA	NA	22.95	24.22	24.73	24.57	24.74	24.58

Table 2. Performance comparison on PSNR values among fixed multi-way search methods

Embedding bits	fixed 3-way search		fixed 4-way search		fixed 5-way search	
	Lena	F16	Lena	F16	Lena	F16
8K	32.01	31.46	31.98	31.45	31.98	31.44
16K	31.41	31.11	31.40	31.12	31.40	31.24
32K	30.27	30.17	30.28	30.17	30.29	30.17
48K	28.67	28.17	28.63	28.19	28.65	28.19
64K	26.55	26.38	26.53	26.38	26.54	26.41
80K	24.73	24.57	24.75	24.57	24.72	24.58

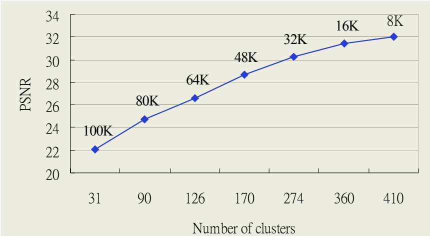


Fig. 3. Relation between the number of clusters and the PSNR values

4 Discussions and Conclusions

Hiding data in VQ compressed images can be easily perceptible since changing a codeword index of the compressed image to a new index value may cause great distortion. In the article, we have presented two data hiding methods, binary search and multi-way search, to make embedding data into the VQ compression image feasible and practical. In addition, the multi-way search scheme we have also brought up can be regarded as the generalized version of binary search.

Instead of applying hierarchical clustering like in Du et al's method, the proposed methods adopt genetic clustering, because hierarchical clustering may lead to worse results when splitting or merging decisions are not well made. On the contrary, genetic clustering guarantees to approach optimal clusters in spherical shapes. The better clusters we get, the better stego-image quality it will be.

The experimental results indicate that the stego-image quality the proposed methods provide is better than that produced by Du et al's method. The results also show that Du et al's method is highly sensitive to the clustering results. To sum up, the proposed methods are suitable for embedding large size secret data without seriously degrading the image quality.

References

1. R. J. Anderson and F. A. P. Petitcolas, "On the limits of steganography," *IEEE Journal on Selected Areas in Communications*, vol. 16, (1998) 474-481.
2. S. Bandyopadhyay and U. Maulik, "Nonparametric genetic clustering: comparison of validity indices," *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, vol. 31, no. 1, (2001) 120-125.

3. W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Systems Journal*, vol. 35, no. 3&4, (1996)313-336.
4. C. C. Chang, D. C. Lin, and T. S. Chen, "An improved VQ codebook search algorithm using principal component analysis," *Journal of Visual Communication and Image Representation*, vol. 8, no. 1, (1997) 27-37.
5. C. C. Chang and H. W. Tseng, "A steganographic method for digital images using side-match," *Pattern Recognition Letters*, vol. 25, no. 12, (2004) 1431-1437.
6. W. C. Du and W. J. Hsu, "Adaptive data hiding based on VQ compressed images," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 150, no. 4, (2003) 233-238.
7. A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*: Kluwer Academic Publishers (1992).
8. R. M. Gray, "Vector Quantization," in *IEEE ASSP Magazine*, (1984) 4-29.
9. S. Katzenbeisser and F. A. P. Petitcolas, *Information hiding techniques for steganography and digital watermarking*: Artech House (2000).
10. R. C. T. Lee, Y. H. Chin, and S. C. Chang, "Application of Principal Component Analysis to Multikey Searching," *IEEE Transactions on Software Engineering*, vol. SE-2, no. 3, (1976) 185-193.
11. Y. C. Lin and C. C. Wang, "Digital images watermarking by vector quantization," *National Computer Symposium*, vol. 3, (1999) 76-87.
12. Y. Linde, A. Buzo, and R. M. Gary, "An Algorithm for Vector Quantization Design," *IEEE Transactions on Communications*, vol. 28, (1980) 84-95.
13. Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, 3 ed: Springer Verlag (1996).
14. M. Ujjwal and B. Sanghamitra, "Genetic Algorithm-Based Clustering Technique," *Pattern Recognition*, vol. 33, no. 9, (2000) 1455-1465.
15. R. Z. Wang, C. F. Lin, and J. C. Lin, "Image hiding by optimal LSB substitution and genetic algorithm," *Pattern Recognition*, vol. 34, no. 3, (2001) 671-683.

Securing Mobile Agents Control Flow Using Opaque Predicates

Anirban Majumdar and Clark Thomborson

Department of Computer Science, The University of Auckland
Private Bag 92019, Auckland, New Zealand
{anirban,cthombor}@cs.auckland.ac.nz

Abstract. Mobile agent technology is an evolving paradigm that combines the inherent characteristics of intelligent agents, namely, adaptability, reactivity and autonomy with mobility. These characteristics of mobile agents provide an excellent means of meeting the distributed and heterogeneous requirements for many electronic commerce applications involving low bandwidth and intermittently connected networks. However, the lack of security in the form of code confidentiality renders this paradigm unsuitable for commercial software. In this paper, we address the problem of mobile agent security by proposing a novel method of mobile agent obfuscation using the concept of opaque predicates to prevent adversaries from observing the control flow of agent code. We discuss about the efficiency of our proposed methodology by demonstrating that to an adversary, the problem of determining the outcome of such opaque predicates is often intractable.

1 Introduction

In the past few years, mobile agent systems have been brought up with the research and development of distributed computing. However, in spite of its tremendous potential, several technical requirements must be met in order to support the widespread transition of agent technology to the commercial domain. Confidentiality of agent code is of foremost concern. The countermeasures directed toward agent protection are radically different from those used for host protection. Host protection mechanisms are a direct evolution of traditional mechanisms employed by trusted hosts and traditional mechanisms are not devised to address threats originating on agents from the execution environment. Agents executing in electronic commerce applications cannot trust the platforms they are executing on and this problem stems from the inability to effectively extend the trusted environment of an agent's host platform to other agent platforms visited by the agent.

Previous works on provable mobile agent security have proposed cryptographic techniques [2] to protect agents from unauthorised code interception; however, since any information belonging to a mobile agent is completely available to its host system, it cannot possibly keep the cryptographic key secret from the system on which it is running. Moreover, the encrypted agents will become susceptible to attacks by the platform once they are decrypted into executable forms. In this paper, we address the problem of agent security using obfuscation [1], which is a technique to obscure the agent code in such a way that an adversary will not be able to gain a complete understanding of its function (with respect to specification and data).

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1065–1071, 2005.

Our work focuses on obfuscating agent behaviours by introducing opaque predicates to guard the control flow. We show that for an adversary to detect the outcome of such predicates, static analyses of the agent code have to be successfully performed and this problem is often intractable in the presence of aliased pointers and concurrency. We also outline the typical scenarios in which such pointer analyses will be difficult to perform. Our technique can be used in conjunction to other obfuscation techniques using aliasing.

2 Mobile Agent Protection Issues and Related Work

The definition of what constitutes an attack depends on what assurances the agent owner needs in order to use a mobile agent. Hohl [3] classified different attack categories that could be mounted on mobile agents by adversaries. We illustrate Hohl's analysis as an attack tree in figure 1. Using our model of protection, we specifically address attacks originating out of spying the control-flow. We achieve this branch confidentiality by obscuring the real control flow of behaviours behind irrelevant statements that do not contribute to the actual computations such that it is impossible for an adversary to find out the correct behaviour from mobile agent code by statically analysing it. An adversary with no semantic understanding of correct control-flow of the code will also find it hard to do purposeful manipulation of the code.

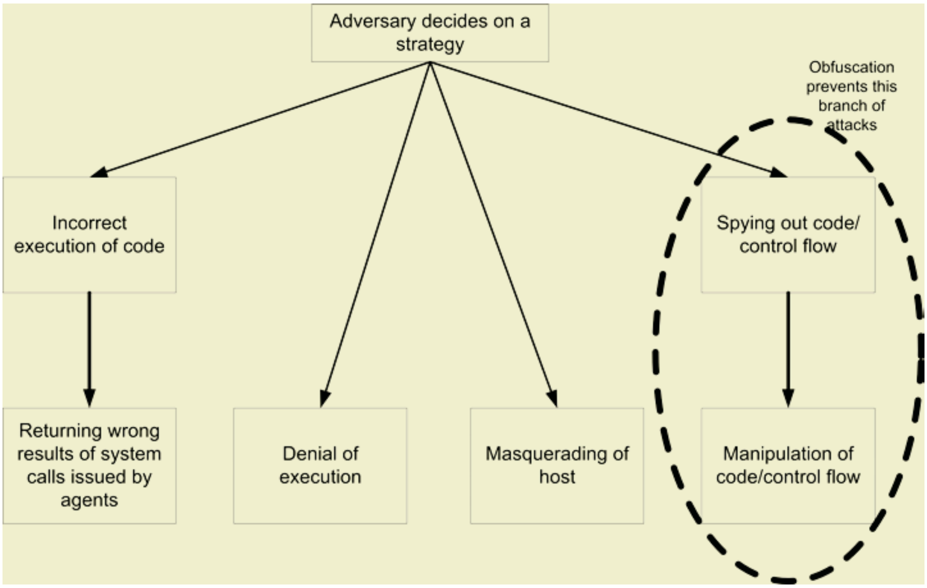

Adversary decides on a strategy
Obfuscation prevents this branch of attacks
Incorrect execution of code
Spying out code/ control flow
Returning wrong results of system calls issued by agents
Denial of execution
Masquerading of host
Manipulation of code/control flow

Fig. 1. The "attack tree". Obfuscation using opaque predicates will attempt to prevent attacks marked with the dotted oval from taking place

Obfuscation for mobile agent code protection was first addressed by Sanders et al [2] in the form of '*mobile cryptography*'. This technique facilitated development of programs that could operate on encrypted data. However, [3][4] points out that this method is not applicable to generic agent codes since it has the restriction that agents can send cleartext data to only trusted hosts. Hohl [3] extended the concept of "black-box security" by incorporating time-limitedness.

However, his method makes explicit assumption of synchronised global clock for token passing between untrusted servers and is therefore difficult to apply in mobile agent interaction scenario which is inherently distributed in nature. Moreover, his technique failed to correlate between the obfuscation techniques and the corresponding time-limitedness they guaranteed. Sakabe et al's [4] attempt to obfuscate mobile agents using aliasing [6][7] is the only work that provides a theoretical basis for obfuscating mobile agents. Wang et al. [5] first proposed an obfuscation technique based on the difficulty of statically analysing aliased pointers in C programs. They manipulated branch targets using aliased pointers and established that the problem of precisely determining indirect branch targets is NP-hard. Sakabe's obfuscation technique takes advantage of polymorphism and exception handling mechanism of Java and established that the problem of *precisely* determining if an instance of a class points-to an overloaded method during an execution of a program is NP-hard.

3 Use of Opaque Predicates for Mobile Agents Obfuscation

Obfuscation is the technique of transforming a program into a form that is more difficult to understand for either a human adversary or for an automated one or both, depending upon the transformation applied [1]. An obfuscated program should have "*identical*" behaviours with respect to the original unobfuscated one. However, we relax this stringent form of restriction by allowing the obfuscated program to have side effects. In this section, we focus on a particular obfuscation class which obscures the control-flow of a program using *opaque predicates.*

An opaque predicate is a conditional expression whose value is known to the obfuscator, but is difficult for the adversary to deduce. A predicate P is defined to be *opaque* at a certain program point *p* if its outcome is only known at obfuscation time [1]. We write $P^F{}_p$ ($P^T{}_p$) if predicate P always evaluates to False (True) at program point *p*. The opaqueness of such predicates determines the resilience of control-flow transformations.

Mobile agent frameworks are instances of loosely-coupled message-passing distributed systems which have the property that communications incur latency and computations proceed at different speeds. Thus, agents do not have any predetermined scheduling policy. This intrinsic nature of mobile agents facilitates programmers to incorporate a large amount of concurrency between them.

If the control-flow of agent P is to be obfuscated using opaque predicates, a certain number of *guard* agents belonging to the system are employed to achieve this task. Since agents in mobile agent systems typically collaborate through message exchanges to achieve a particular task, the set of guard agents could be those that agent P frequently communicates with. The actual number of guard agents employed in protection of a single agent will depend *dynamically* on the availability of agents in the system. It then initialises a data structure such as circular linked-list with some initial pointers pointing on it. Agent P then sends a message, containing this data structure, to one of the guards. Each guard, when it receives the data structure, may change one of its pointers and will then try to send the data structure along to another guard or back to agent P. The message-passing pattern and pointer update by guards must maintain an invariant that holds on the data structure when it is sent back to P. We shall illustrate this protocol by a simple example as depicted in figure 2. Let four guard agents be dynamically spawned by P. We call these guards A, B, C, and D. Agent P initialises a circular linked-list and passes it randomly to a guard which initiates the message-passing between itself and other guards. Let the A be the initial guard agent and it updates pointer *p*. Similarly, guards B, C, and D manipulate pointers *q*, *r*, and *s* respectively.

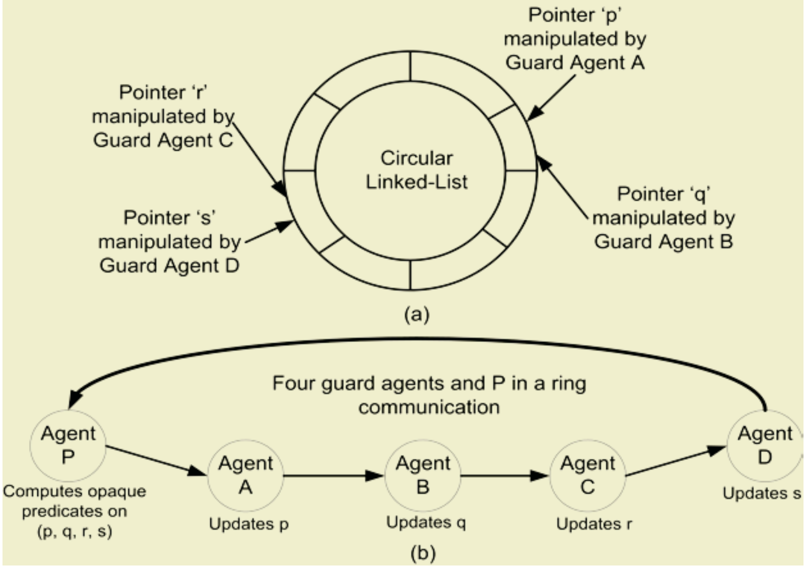

Pointer 'p' manipulated by Guard Agent A
Pointer 'r' manipulated by Guard Agent C
Circular Linked-List
Pointer 'q' manipulated by Guard Agent B
Pointer 's' manipulated by Guard Agent D
(a)
Four guard agents and P in a ring communication
Agent P
Computes opaque predicates on (p, q, r, s)
Agent A
Updates p
Agent B
Updates q
Agent C
Updates r
Agent D
Updates s
(b)

Fig. 2. The global dynamic data structure in the form of a circular linked-list shared by four guard agents A, B, C, and D and a simple pointer update protocol

During the initialisation process, P also embeds the linked-list update-invariants in each of the guard agents such as:

- *p* and *q* are aliased.
- *r* and *s* are aliased.
- *p* and *q* never alias pointers *r* and *s* and vice-versa.

In figure 2(b), we have illustrated the communication pattern between agent P and guards with a simple ring protocol. In practical systems, however, more complicated communication patterns could be used. After updating its respective pointer, each guard agent passes on the data structure to its successor. Finally, agent P receives the data structure from guard D, after which, it proceeds to construct and check opaque predicates using aliased pointers *p*, *q*, *r*, and *s* as illustrated in the pseudo-code snippet of figure 3.

The branch obfuscated using opaque predicate [(p==q) && (r==s)] only evaluates to true; whereas, branches obfuscated using any other combination evaluate to false. Before checking for the outcome of opaque predicates, the obfuscated agent P waits for the data structure from guard agent D. After receiving the data structure, it then sends it around for another pointer update round. These data structure passing messages could be tagged with a special value to distinguish from other messages.

```
Guard Agent:
      …
       //initialisation
       receive <update_rule, Agent(P)>
       //perform update
       while (true) {
             receive <data_structure, Agent(ID)>;
             update (pointer);
             send <data_structure, Agent(ID)>;
      …
       }

Obfuscated Agent P:
       …
       initialise (data_structure);
       send <data_structure, Agent(A)>;
       …

       //receive the data structure from Guard Agent D
       while (!receive <data_structure, Agent(D)>) {
             wait;
      }
       //send the data structure for another round of updates
       send <data_structure, Agent(A)>;
       //initiate testing on pointer invariants

       if    (data_structure.p    ==    data_structure.r    &&
data_structure.q == data_structure.s)
       { // Opaquely-False Predicate
             // perform dummy behaviour
       }
       else {
             //perform real behaviour
       }
       if    (data_structure.p    ==    data_structure.q    &&
data_structure.r == data_structure.s)
       { // Opaquely-True Predicate
             // perform real behaviour
       }
       else {
             //perform dummy behaviour
       }
```

Fig. 3. Pseudo-code showing sample guard agent pointer update action and obfuscation of control-flow in P using the aliased opaque predicates

In order to statically analyse the obfuscated code, the adversary must depend on a static slicer to find those parts of a program which could affect the value of opaque predicates at points of interest (namely, at locations where behaviours are obfuscated). The slicing of distributed programs is a major challenge due to the timing related interdependencies among processes. Moreover, to find the slicing criterion of the slicer, the analyser must rely on alias analysis [6][7][8] to determine what kind of structure the pointers point to at runtime, and if the pointer corresponding to the variables used in the construction of opaque predicate may/must refer to the same dynamic object in the guard agents at some program location (where the opaque predicates are used in P). Two pointers referencing the same memory location are called *aliases*. Therefore, the static analyser must use inter-process escape alias analysis to determine the objects that can be referenced by pointers in processes other than the processes in which they are allocated.

Though numerous works on intra- as well as inter-procedural alias analysis and inter-procedural thread escape analysis [9][10] have been done throughout the last few decades, we have not come across a method that can perform alias analysis by considering asynchronous message-passing of distributed processes as escape points. We believe the reason why this problem has not yet been addressed by the program analysis community is because we still do not have efficient, precise and scalable algorithms for performing simpler cases of alias analysis in sequential multi-threaded programs and hence are unsure how to go about solving this problem which is much more complex in nature.

4 Conclusion

In this paper, we have addressed the problem of mobile agent code obfuscation using opaque predicates, which are structures inserted at program control points to obfuscate the branching of agent behaviours. Obfuscation using opaque predicates that use the inherent concurrency associated with mobile agent systems and aliasing are resilient against well known static analysis attacks. We have demonstrated that it will be very difficult for an adversary to understand the dynamic structure of the predicate and mount attacks that could statically analyse the associated values for each of the terms present in the predicate. The predicate structure can also be made arbitrarily complex by incorporating numerous guard agents dynamically and this demonstrates the flexibility of our technique. Moreover, in an already existing message-passing scenario between agents, the messages exchanged between the obfuscated agents and the guards will contribute to a negligible amount of extra overhead. The technique also does not depend of any particular language feature and is therefore applicable to generic mobile agent platforms.

We note that our technique is not an alternative to the one proposed by Sakabe et al. Obfuscation using method aliasing as a standalone technique may not be sufficiently resilient since the smaller size of agent programs compared to that of commercial Java software may result in a fewer number of methods to be considered for overloading in Sakabe's technique. Hence, using our technique in conjunction to the one already proposed by Sakabe et al will substantially strengthen the resilience of obfuscated agents.

In this contribution, we make the assumption that an adversary is only able to statically manipulate an agent executing on a platform but not its guards. By imposing this restriction, we prevent the adversary from getting a complete understanding of data structure passing and pointer update rules. Future work will be concentrated on investigating the classes of resilient opaque predicates that demonstrate provable resistance against well known static analysis attacks as well as dynamic analysis attacks including dynamic message interception attacks.

References

1. Collberg, C., Thomborson, C., and Low, D.: *Manufacturing Cheap, Resilient, and Stealthy Opaque Constructs*. In Proceedings of 1998 ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL'98). 1998.
2. Sander, T., and Tschudin, C.F.: *Protecting mobile agents against malicious hosts.* In Vigna G., ed.: Mobile Agents and Security. Volume 1419 of Lecture Notes in Computer Science. Springer-Verlag. 1998
3. Hohl, F.: *Time limited blackbox security: Protecting mobile agents from malicious hosts.* In Vigna G., ed.: Mobile Agents and Security. Volume 1419 of Lecture Notes in Computer Science. Springer-Verlag. 1998
4. Sakabe, Y., Masakazu, S., and Miyaji, A.: *Java obfuscation with a theoretical basis for building secure mobile agents*: . In Lioy A., Mazzocchi D. eds.: Communications and Multimedia Security (CMS 2003). Volume 2828 of Lecture Notes in Computer Science. Springer-Verlag. 2003.
5. Wang, C., Hill, J., Knight, J.C., and Davidson, J.W.: *Protection of software-based survivability mechanisms*. In Proceedings of the 2001 conference on Dependable Systems and Networks. IEEE Computer Society. 2001.
6. Horwitz, S.: *Precise Flow-insensitive may-alias in NP-hard*. ACM Transactions on Programming Languages and Systems (TOPLAS), Vol. 19 No. 1, 1997.
7. Hind, M., Burke, M., Carini, P., and Choi, J.D.: *Interprocedural pointer alias analysis*. ACM Transactions on Programming Languages and Systems (TOPLAS), Vol. 21 No. 4, 1999.
8. Rugina, R. and Rinard, M.: *Pointer analysis for multithreaded programs*. In Proceedings of 1999 ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI '99). Atlanta, GA, USA. 1999.
9. Salcianu, A. and Rinard, M.: *Pointer and escape analysis for multithreaded programs*. In Proceedings of the 2001 ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPOPP '01), Snowbird, UT, USA. 2001.
10. Whaley, J. and Rinard, M.: *Compositional pointer and escape analysis for Java programs*. Proceedings of the 1999 ACM SIGPLAN Conference on Object-Oriented Programming Systems, Languages & Applications (OOPSLA '99), Denver, CO, USA. 1999.

A Verifiable Fingerprint Vault Scheme

Qiong Li, Xiamu Niu, Zhifang Wang, Yuhua Jiao, and Sheng-He Sun

Information Countermeasure Technique Institute
Harbin Institute of Technology, Harbin, Heilongjiang, China
{qiong.li,xiamu.niu}@dsp.hit.edu.cn

Abstract. By adopting a non-interactive information-theoretic secure verifiable secret sharing scheme in an unorthodox way, a verifiable fingerprint vault scheme is presented in this paper. Fuzzy vault scheme is a novel cryptographic construct which can increase the security of the biometric template in a biometric authentication system. It can be also used to bind the cryptographic key and the user in a cryptosystem to overcome the security problems caused by the key stolen, key illegally shared, etc. The main weakness of the scheme is the computation complexity is too high for the legitimate user to unlock the vault. In this paper, a verifiable fingerprint vault is proposed to improve the fuzzy vault scheme. It is shown that our scheme is effective.

1 Introduction

Although the biometric authentication has become a popular identity authentication solution in the field of information security, it is still facing some security problems. One of these problems is the storage of the biometric templates that does not exist in passwords-based authentication [1]. As we know, in many implementation of UNIX, a given password *P* is not stored explicitly in the system. Instead, it is stored in the form of hash digest of the password and a salt as *Hash(P, salt)* [2]. Thanks to the features of collision free and onewayness of a good hash function, the hash digest can be stored in a public directory without comprising security. Unfortunately, such effective solution cannot be adopted to store the biometric template due to the variances existed in the biometric authentication. Take fingerprint as an example, two scans of the same finger rarely lead to the identical fingerprint images. Consequently, the fingerprint process algorithm will generate different fingerprint templates for the different fingerprint images. To handle the inherent variances in biometric authentication, most systems store a template for each user in explicit form. Yet the protection of biometric data is necessary since the user's biometric data is the representation of his identity, besides, biometric data is non-revocable [3]. The "fuzzy vault scheme" (FVS) proposed by Ari Juels and Madhu Sudan can be a solution to improve the security of the biometric template [4]. In FVS, a long and strong key is locked by the user's fingerprint in a "vault". Only a matching but not necessary identical fingerprint can unlock the vault and retrieve the accurate key. Meanwhile, it is computation unfeasibly for the attacker to unlock the vault. The key protected by the user's fingerprint can be used as any cryptographic key, so FVS can also solve the problem of the lack of connection between the cryptographic key and the user, which exists in the traditional cryptosystems [5].

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1072–1078, 2005.

But it is found that the computation complexity for the valid user to unlock the vault in FVS is still very high, and FVS dose not provide a mechanism to check if the unlocking is success or not. By combining a non-interactive information-theoretic secure verifiable secret sharing scheme, a verifiable fingerprint vault scheme (VFVS) is presented in this paper. The VFVS can decrease the valid user's computation complexity and provide a method to judge the operation result.

2 Review and Analysis of FVS

In this section, the construction of FVS is introduced briefly and its weakness is analyzed.

2.1 FVS

FVS combines fingerprint authentication and polynomial interpolation to lock a secret $\mathcal{K}$ in a vault by the user's fingerprint template and unlock the vault by a matching fingerprint which is close to the template but not necessary identical. At first, a polynomial $f(x)$ of degree less than k over a finite field F is selected such that $f(x)$ encodes the secret $\mathcal{K}$ in some way (e.g., has an embedding of $\mathcal{K}$ in its coefficients). A user's registration fingerprint template, e.g. a minutia set $fp = \{m_i\}_{i=1}^{t}$ is projected onto the polynomial and the pairs $\{m_i, f(m_i)\}_{i=1}^{t}$ are stored along with some noise points $\{c_i, d_i\}_{i=t+1}^{r}$ ($c_i \neq m_i$, $d_i \neq f(c_i)$) such that $fp = \{m_i\}_{i=1}^{t}$ is concealed by these noise points. The pairs $\{m_i, f(m_i)\}_{i=1}^{t}$ and $\{c_i, d_i\}_{i=t+1}^{r}$ constitute the vault V .

Given the matching fingerprint fp' of the valid user, which normally overlaps fp to some extent enough true points can be separated from V to construct the unlocking set U . As a result, U can be used to reconstruct the polynomial and retrieve the secret $\mathcal{K}$. Lemma 1 proves that with high probability, many polynomials of degree less than k agree with t elements in V . In another words, the true polynomial is concealed by many spurious polynomials. The illegal user or attacker cannot distinguish the noise and the true points. Therefore, neither can they pick the true polynomial. Please refer to [4] for the proof of Lemma 1.

Lemma 1. For $\forall \mu > 0$, with probability at least $1-\mu$, there exist at least $\frac{\mu}{3} q^{k-t} (r/t)^t$ polynomial $f'(x)$ of degree less than k such that V include exactly t elements of the form $(x, f'(x)) \in F \times F$.

2.2 Analysis of FVS

From the description as above, it can be found that the security of FVS depends on the number of noises in V . The more noise points are added into the vault, the greater probability that some set of these noise points and/or true points align themselves on some polynomial of the desired degree. But more noise not only means the vault is harder to be broken by an attacker, but also harder to be unlock for the valid user.

In FVS, every minutia is represented by its coordinates in the Euclidean plane. Suppose that when the Euclidean distance between two minutiae is less than δ, i.e. $|m_i, m_i'| < \delta$, two minutiae are matching. When there are at least *th* matching minutiae in fp' and fp, two fingerprint templates are matching. There are many reasons for fingerprint template variations, such as the finger's displacement, rotation, nonlinear distortion, pressure, skin condition and feature extraction errors, etc [6]. So it is almost impossible to have two identical fingerprint templates for one finger's different scans. When $r > t$, $|c_i, m_i'| < \delta$ may happen when a valid user tries to unlock the vault using a new sampled fingerprint. So a valid user's unlocking set U may include some noise points. Let us suppose that during the unlocking operation, U is consisted of *t'* elements selected from *V* as $U = \{m_i, f(m_i)\}_{i=1}^{s} + \{c_i, d_i\}_{i=s+1}^{n}$. As long as $s > k$, the user can reconstruct the polynomial by $\mathrm{C}_n^s / 2$ polynomial interpolation trials on an average. In fingerprint authentication, a fingerprint template normally includes about 40 minutiae and two fingerprint templates match when there are more than 12 minutiae match [7]. Suppose that the user separate 30 elements from *V* and 18 of them are true, i.e.*n*=30, *s*=18, the average computation is 4.3247e+007 trials. Such high computation complexity is unacceptable in a real-time application. Another problem of FVS is the scheme does not provide a mechanism to check if and when the unlocking is successful.

3 Verifiable Fingerprint Vault Scheme

In order to overcome the problems mentioned above, a verifiable fingerprint vault scheme is proposed in this paper based on the non-interactive information-theoretic secure verifiable secret sharing scheme proposed by Pedersen in [8].

3.1 Pedersen 's Verifiable Secret Sharing Scheme

The secret sharing scheme (SSS), also called threshold scheme and key safeguarding scheme, was introduced by Shamir and Blakeley independently in 1979[9,10]. Since then, SSS has been playing an important role in the filed of information security. In a (*n*, *k*) SSS scheme, the dealer divides a secret *x* into *n* shadows $y_1, \ldots y_\mathrm{n}$ and sends one shadow to each player secretly. The original secret can be recovered by any *k* valid shadows but cannot be determined by any *k*-1 or less valid shadows. In order to deal with the possible deception among the dealer and players, verifiable secret sharing scheme (VSS) has been studied to verify if a player has received a shadow consistent with the other shadows since 1985 [11]. Pedersen produced the first non-interactive information-theoretic secure VSS [8]. Pedersen's VSS was built for Shamir's polynomial interpolation SSS. The VSS takes advantage of a homomorphic commitment scheme to generate the verification data for each shadow during the secret distribution. In the homomorphic commitment scheme, g is the generator of G_q, $\forall h \in G_q$. It is obvious that the computing of $\log_g h$ is a hard problem, i.e. the computing of discrete logarithm in a finite field. The committer commits himself to an $s \in Z_q$ by choosing $t \in Z_q$ at random and computing

$$E(s,t) = g^s h^t . \qquad (1)$$

Such a commitment can later be opened by revealing *s* and *t*. It has been proved that $E(s,t)$. reveals no information about *s*, and the committer cannot open a commitment to *s* as $s' \neq s$ unless he can find $\log_g h$. So the verification data in Pedersen's VSS reveals no information about the shadow either.

3.2 Locking and Unlocking Algorithms of VFVS

It is shown that the threshold is important for both SSS and FVS. The threshold of SSS decides if the original secret can be recovered; the threshold of FVS decides if the true polynomial can be reconstructed. Actually, FVS can be considered as an application of Shamir's SSS. Similarly, the noise points in the unlocking set U can be considered as the fake shadows due to some players' deception. VSS detects the deception through the pre-broadcasted verification data. In our VFVS, we take advantage of the verification data to check the validity of elements in the unlocking set. By removing the noise pointes before the polynomial reconstruction, the valid user's computation complexity can be decreased greatly. VFVS include the Locking and Unlocking algorithms as following.

Locking Algorithm:

1. Over the finite field Z_q, connect the secret κ with a polynomial $f(x)$ of degree less than k $f(x) = a_0 + a_1 x + ... + a_{k-1} x^{k-1}$, construct another a polynomial $g(x)$ of degree less than k $g(x) = b_0 + b_1 x + ... + b_{k-1} x^{k-1}$
2. Project a user's registration fingerprint template $fp = \{m_i\}_{i=1}^{t}$ onto the two polynomial $f(x)$ and $g(x)$ to form pairs $(m_i, \ \alpha_i = f(m_i))$, $(m_i, \beta_i = g(m_i))$, the locking set is $L = \{m_i, \alpha_i, \beta_i\}_{i=1}^{t}$
3. Compute $E_i = g^{\alpha_i} h^{\beta_i}$, $0 \le i \le t$, and keep it with the user
4. Randomly select noise points $(c_i, d1_i), (c_i, d2_i)$, such that $c_i \neq m_i$, $d1_i \neq f(c_i)$, $d2_i \neq g(c_i)$, unite these noise points and the locking set L to constitute the vault $V = \{m_i, \alpha_i, \beta_i\}_{i=1}^{t} \bigcup \{c_i, d1_i, d2_i\}_{i=t+1}^{r} = \{Lx_i, Ly1_i, Ly2_i\}_{i=1}^{r}$
5. Sort the vault V in descending sequence by Lx_i and store it at the client-side computer.

Unlocking Algorithm:

1. Given the user's authentication fingerprint $fp' = \{m_i'\}_{i=1}^{n}$, compare every m_i' with the elements of V, if $(m_i' - Lx_j) \le \delta / \sqrt{2}$, add $(Lx_j, Ly1_j, Ly2_j)$ into the unlocking set U
2. For every element in U, check if the equation $E(Ly1_j, Ly2_j) = \prod_{i=0}^{k-1} (E_i)^{Lx_j^{\,i}}$ holds

3. If there are at lest k elements in U pass the verification in the step 2, the polynomial $f(x)$ can be reconstructed by these verified elements successfully, otherwise the unlocking fails.

The verification of each element in U is based on the homomorphism of the commitment (1). If the element is a true element, then the equation (2) holds.

$$\begin{aligned} E\left(Ly1_j, Ly2_j\right) &= g^{Ly1_j} h^{Ly2_j} = g^{f\left(Lx_j\right)} h^{g\left(Lx_j\right)} = g^{a_0 + a_1 Lx_j + \ldots + a_{k-1} Lx_j^{k-1}} h^{b_0 + b_1 Lx_j + \ldots + b_{k-1} Lx_j^{k-1}} \\ &= g^{a_0} h^{b_0} \bullet (g^{a_1} h^{b_1})^{Lx_j} \cdots (g^{a_{k-1}} h^{b_{k-1}})^{Lx_j^{k-1}} = E_0 \bullet E_1^{Lx_j} \cdots (E_{k-1})^{Lx_j^{k-1}} = \prod_{i=0}^{k-1} (E_i)^{Lx_j^i} . \end{aligned} \tag{2}$$

3.3 Computation Complexity Comparison Between FVS and VFVS

As mentioned above, the average computation complexity of a valid user in FVS is $C_n^s / 2$ polynomial interpolation trials. Take the commonly used Lagrange interpolation algorithm into consideration, the interpolation formula is listed in (3)

$$l\ (x) = \sum_{i=0}^{k-1} l(x_i) P_i(x) \tag{3}$$

where $l(x_i)$ are the known values of the function and $l(x)$ is the desired polynomial of degree k-1. The Lagrange polynomial $P_i(x)$ is the polynomial of order k-1 that has the value 1 when $x = x_i$ and 0 for all $x_{j \neq i}$:

$$P_i(x) = \prod_{\substack{i=0 \\ j \neq i}}^{k-1} \frac{(x - x_j)}{(x_i - x_j)} \tag{4}$$

According to the equation (3) and (4), one Lagrange interpolation needs about $(2k^2 - k)$ multiplication operations to get a polynomial of degree k-1. So the unlocking operation is about $C_n^s (k^2 - k/2)$ multiplications.

In our VFVS, the computation consists of the verification and one polynomial interpolation. Since the commitment as (1) can be done by less than $2|q|$ multiplications [7], the verification of n elements in U can be done by less than $(2|q| + |q|k)n$ multiplications, where $|q|$ indicates the length of q in binary representation. So the total computation is about $(2|q| + |q|k)n + 2k^2 - k$ multiplications. Let us compare the computation complexity between the two schemes when these parameters are assigned reasonable values: $n = 30, s = 18,\ k = 12, q = 256$. The computation of FVS is about 3e+006 times of that of VFVS.

4 Conclusions

A verifiable fingerprint vault scheme is proposed in this paper to improve the computation efficiency of Juels and Sudan's fuzzy vault scheme. By adopting Pedersen's non-interactive information-theoretic secure verifiable secret sharing scheme, our scheme can simplify the unlocking operation of the valid user and make it more practical in real applications. It is shown that the improvement is encouraging. The author's future work will be the experiment of our scheme in a real fingerprint authentication system.

Acknowledgment

This work is supported by the Foundation for the Author of National Excellent Doctoral Dissertation of P. R. China FANEDD-200238, the National Natural Science Foundation of China 60372052, the Multidiscipline Scientific Research Foundation of Harbin Institute of Technology HIT.MD-2002.11, Program for New Century Excellent Talents in University, and the Foundation for the Excellent Youth of Heilongjiang Province.

References

1. Marcos Faundez-Zanuy, "On the vulnerability of biometric security systems", IEEE A&E systems magazine, June 2004, p3-8
2. D.C.Feldmeier, P.R. Karn. "UNIX password security-ten years later". In G.Brassard, editor, Advances in Cryptology – CRYPTO'89, page 44-63. Springer-Verlag, 1989. LNCS No.435
3. A. Juels, M.Watenberg, "A Fuzzy Commitment Scheme", Proc. 6th ACM Conf. Computer and Communications Security, 1999, pp. 28–36
4. Juels, A., "A fuzzy Vault Scheme", ACM Conference on Computer and Communications Security, CCS2002
5. U.Uldaag,, S.Pankanti, "Biometric Cryptosystems: Issues and Challenges", Proceeding of the IEEE, 2004, 92 (6), pp.948-960
6. D. Maltoni, D. Maio, A. K. Jain, S. Prabhakar, "Handbook of Fingerprint Recognition". New York: Springer-Verlag, 2003
7. T.P. Pedersen, "Non-interactive and Information-Theoretic Secure Verifiable Secret Sharing [A]", CRYPTO'91 [C], 1991
8. A. Shamir, "How to Share a Secret", Communications of the ACM, Vol 22, 1979, pp. 612-613
9. G.R. Blakley, "Safeguarding Cryptographic Keys", Proceedings of the National Computer Conference [C], Montvale: NCC, 1979
10. B. Chor, S. Goldwasser, S. Micali, and B. Awerbuch, "Verifiable Secret Sharing and Achieving Simultaneity in The Presence of Faults [A]", Proceedings of 26th IEEEE symposium on Foundations of Computer Science, 1985
11. X.B. Cai, B.Z.Ceng, "Civil Fingerprint Recognition Technique". Posts and Telecom Press, 2004

Appendix

In FVS, Reed Solomon decoding algorithm *RSdecode* is suggested to be used to improve the computation performance: *RSdecode* takes as input a collection of points which are presumed to lie preponderantly on a single polynomial of pre-specified degree at most $k-1$; if successful, *RSdecode* output a polynomial $f(x)$, intersecting the majority of input points. The author chose the classical RS decoding algorithm of Peterson-Berlekamp-Massey (PBM). Although the encoding experts have developed several Interpolation-based RS decoding algorithms, such as Guruswami-Sudan algorithm, the application of RS decoding algorithm of PBM in polynomial interpolation has not been found so far. A short paragraph of Matlab code shows how RS decoding algorithm of PBM works.

```
m = 3; % Number of bits per symbol
n = 2^m-1; k = 3; % Word lengths for code
msg = gf([2 7 3; 4 0 6; 5 1 1],m); % Three rows of m-bit symbols
code = rsenc(msg,n,k);
errors = gf([2 0 0 0 0 0 0; 3 4 0 0 0 0 0; 5 6 7 0 0 0 0],m);
noisycode = code + errors;
[dec,cnumerr] = rsdec(noisycode,n,k)


dec = GF(2^3) array. Primitive polynomial = D^3+D+1 (11 decimal)
Array elements =[2    7    3;    4    0    6;    4    0    0]
cnumerr =  [1     2    -1]'
```

The Research on Information Hiding Based on Command Sequence of FTP Protocol

Xin-guang Zou[1], Qiong Li[1], Sheng-He Sun[1], and Xiamu Niu[1]

[1] Information Countermeasure Research Institute, Harbin Institute of Technology
P.R.C. 150001
xgzou@dsp.hit.edu.cn

Abstract. Traditional covert channels study mainly focus on digital signature, operation system and multimedia technology. The covert channels in TCP/IP protocol suite have many advantages, such as they are harder to be detected and more robust. Two covert channels are discovered by using FTP protocol command sequence based on the study of FTP protocol. The principles of covert channels are analyzed. Three characters of covert channels, which are concealment, bandwidth and robustness, are studied. And the methods to construct covert channels are also applicable to other internet protocol.

1 Introduction

Traditional covert channels study mainly focus on digital signatures, operation system and multimedia technology. It is found that the covert channels based on TCP/IP protocol suite have many advantages. The TCP/IP covert channels are implemented by utilizing reserved or redundant bits in TCP/IP header.

There are many advantages of the TCP/IP covert channels, which make the covert channels harder to be detected and more covert information bits can be sent.

(1) The TCP/IP packets are created and deleted dynamically, which means they are not stored in the network for a long time. Therefore it is hard for the attacker to collect all the packets, which carry the hiding information, to get the covert information correctly.
(2) TCP/IP communication is very common today. Every two computers may have TCP/IP traffic with each other, which makes it securer than traditional covert channels. The channel has good concealment character.
(3) The megabit switch local area network is the mainstream today and gigabit network may evolve in near future. The backbone network has tens of gigabit per second bandwidth now. The TCP/IP covert channels' bandwidths are very rich.
(4) The packets, which carry covert information, can transfer through various routes due to the distribution of the TCP/IP network and dynamic route protocol update. It is harder to the attacker to get the whole covert carries in such highly distributed environment.
(5) The authentication information can be embedded into the TCP/IP covert channels to make the authentication implicit, which will not arouse the attacker's attention.
(6) There are hundreds of application protocols using TCP/IP protocol as the carrier. Every application's implementation is complicated. These two factors make the covert channels based on TCP/IP application protocol even harder to be detected than others.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1079–1085, 2005.

FTP protocol is used to transfer files between clients and servers. Two kinds of cover channels are discovered and studied in the paper. In the following sections, we discuss the related work and study the two kinds of covert channels. In section 2, we introduce the file transfer protocol in brief. In section 3, we introduce the related work on TCP/IP covert channels. In section 4 and 5, we analyze the concealment, bandwidth and robustness of the two kinds of covert channels by using FTP command sequence. The section 6 is our conclusion.

2 File Transfer Protocol

FTP is a TCP/IP application protocol, which implements the file transfers function [1]. It is defined in some IETF request for comment, such as rfc959, etc. FTP generally uses the TCP as the carrier protocol in practical application. As a kind of multiple connections protocol, FTP's connections are divided into control connections and data connections. The control connections adopt Telnet protocol to transmit the commands and responses and the data connections adopt TCP protocol to download and upload the file data.

Total 33 commands defined in rfc959 are divided into three categories, which are access control commands, transfer parameter commands and ftp service commands. The commands discussed in the paper are limited into rfc959 and the extensions of FTP are not considered.

3 Related Works

The current research on TCP/IP covert channels mainly uses IP, TCP, UDP header or ICMP payload as the covert information carrier [2-5]. The proposed algorithms generally make use of the reserved bits, unused bits in special environment and bits, which can be set randomly, to carry the covert information. Due to the simplicity of the protocols, the covert channels are easy to be detected and destroyed.

The 4 classes of proposed covert channels will be introduced in detail.

1. IP identification field covert channel. The length of IP identification field is 16 bits. A byte of covert information can be embedded into the high order 8 bits of identification field and the low order 8 bits are set randomly to avoid producing same values.
2. TCP initial sequence number covert channel. The ISN of TCP is 32 bits, which can transmit 32 bits covert information in one TCP connection.
3. TCP source address covert channels. Every IP packet has a 32 bits source address field, which can carry 32 bits covert information.
4. ICMP echo request/reply covert channel. The payload length of the ICMP packet is various, which can carry a number of information.

The covert channel 1 and 2 are based on random number information hiding, which are good concealment character but are fragile. The third channel does not work when the gateway performs the source address filter. The ping packet payload generated by mainstream operation system is fixed. It is suspicious when the payload is set randomly. The fourth channel is easy to be detected. To destroy the covert channel, a ping proxy can be set up to rewrite the content of ping packet.

As we can see, there are a number of limitations in covert channels using TCP/IP packet header. On the other hand, application layer information is checked and rewritten seldom in current network node. Due to the complicity of application protocol, it is harder to be detected than above channels. FTP protocol discussed in the paper is a kind of application layer protocol used extensively, which has complicated control scheming and huge volume of data transfers. The concealment of FTP covert channels is excellent.

In general, the covert channels should have three important characters, which are concealment, bandwidth and robustness. Concealment is the capability of resisting passive attack. Bandwidth means how many covert bits can be sent in unit time or unit covert carry. Robustness is the ability of resisting active attack.

In the following sections, two kinds of covert channels will be studied. The first kind of covert channels are based on FTP command mapping scheme and the second ones are based on FTP command sequence.

4 The Covert Channels Based on FTP Command Mapping

In file transfer protocol, clients and servers exchange commands and responses to control the file transfer. Every command can be encoded into a bit-string. But in rfc959, the commands are encoded into ASCII format. So, we can build a new commands encoding algorithm, which can use the commands to encode the covert information.

We select a command space Q, which is a set of FTP commands. Every command in the Q can be encoded by a fix length of bits. Suppose the size of Q is N, which means there are N commands in the Q. The number of bits used to represent a command can be computed in the formula 1.

$$\log_2 N \tag{1}$$

We can also adopt the hierarchical encoding method. The command space Q is divided into K command sub spaces Q_1, Q_2,…Q_K, whose sizes are N_1, N_2, …, N_K correspondingly. And the sub-space sizes satisfy $N = \sum_{i=1}^{m} N_i$. Each command in *k*th sub space can be encoded into $\log_2 N_k$ bits, at the same time, the *k*th sub space index can be encoded into $\log_2 K$ bits. So, each command in *k*th sub space's encoding length is defined in formula 2.

$$len_k = \log_2 N_k + \log_2 K \tag{2}$$

Theorem: Let the size of a command space θ be *N*, which satisfies $N = 2^l$ and the number of command sub space to be *K*, which satisfies $K = 2^m$. The sub spaces are named $\theta_1, \theta_2, ..., \theta_K$ correspondingly. The size of θ_i is N_i, which satisfies $N_1 = N_2 = \ldots = N_K$. Then any command in space θ has the same encoding length.

Proof: The size of space θ_i satisfies $N_i = {}^{N}\!/_{K} = {}^{2^l}\!/_{2^m} = 2^{l-m}$. According to formula 1, the size of encoding length of command in space θ_i satisfies the following relationship:

$$len_{\theta_i} = \log_2 N_i = \log_2 2^{l-m} = l - m \tag{3}$$

At the same time, the space θ_i 's index encoding length is

$$len'_{\theta_i} = \log_2 K = \log_2 2^m = m\,. \tag{4}$$

So according to formula 2, the total encoding length of command in sub space θ_i is

$$len_{\theta_i} + len'_{\theta_i} = l - m + m = l \tag{5}$$

It means that the encoding length of command in sub space θ_i is a constant, which has no relationship with the sub space's index. So, the theorem is proved.

According to the theorem, the covert channels making use of command encoding algorithm has the same bandwidth, whenever adopt the whole command space or sub spaces with same size to encoding the covert information.

As the commands sent in the covert channel are decided by the covert information, the covert channel will inherit the statistical characters of the covert information. If the covert information has a prominent statistical distribution character, the covert channel will be recognized easily. To solve the problem, the original covert information must be encrypted to randomize the bitwise probability distribution.

If we select 32 commands from the command set defined in ftc959 to organize a command space, the bandwidth of covert channel is 5 bit per command according to above theorem. The covert channel has strong robustness character, since TCP's retransmission scheme can protect the channel from disturbing noise. Active attacker must cut off the normal FTP communication to eliminate the covert channel because there is no covert information embedded into the packets, which can be destroyed just by rewrite the corresponding bits.

5 The Covert Channels Based on FTP Command Sequence

The common used FTP client software integrates a function called idle prevention scheme, which guarantees at least one command will be sent in a fixed period of time. The user could be allowed to select which commands will be sent to implement the idle prevention scheme. The covert channels based on FTP command sequence are based on the idle prevention schemes. To illustrate the principal of covert channel, we suppose three commands such as NOOP, ABOR and ALLO are selected in idle prevention function and can be sent at will by the user.

NOOP command is used to protect the FTP control connections timeout in the condition of a long time idle. When a server receives NOOP command, the server just resets the connection idle timer and replies "200 Command okay". ABOR command makes the server abort a command, which is executing currently, and close the relative data connections. If no command is executing when a server receives an ABOR, the server just replies "226 ABOR command successful." ALLO command requests a server to reserve some buffers. Some kinds of server just reply a command successful response and do nothing in fact.

The covert channels use a special sequence of these commands to encode a fix length of covert information. In FTP implementation, these commands needed be sent at random to keep the control connections from idle condition.

The kind of covert channel can be divided into two sub-class. The first class of covert channels utilized only two of the three commands to work, which is called sequence length covert channel. The second class of covert channels utilizes the whole three commands to work, which is called hierarchical sequence length covert channel. The schemes of the two covert channels and comparison of the bandwidths between the two channels will be discussed in detail in following sections.

5.1 Covert Channel Based on Sequence Length

The covert channel utilizes the number of sent NOOP commands to encode a byte of the covert information. To transmit a byte of covert information, the number range of sent commands is from zero to 255. Before a byte of covert information is transmitted, an ABOR command needs to be sent to represent the beginning of next byte of covert information.

Suppose the probability distribution of covert information bytes is uniform. To transmit *N* bits covert information, the number of FTP commands need to be sent is

$$\frac{1}{2^N}\sum_{i=0}^{2^N-1}(i+1). \tag{6}$$

When a byte of cover information value is *i*, it means *i* NOOP and one ABOR commands need to be sent to represent the byte of covert information.

The covert channel has good concealment because it is a normal scheme sending NOOP or ABOR continually to prevent the control connections from entering idle status. The huge cost is a shortage of the covert channel. On the average, 128.5 commands need to be sent to transmit a byte of covert information. It is fit for transmitting short message. An improved covert channel will be discussed in the next section, which has better encoding performance.

5.2 Covert Channel Based on Hierarchical Sequence Length

The covert channel utilizes NOOP, ABOR and ALLO commands sequence to perform hierarchical encoding. Divide *N* bits covert information into two parts: high order *M* bits and next low order *N-M* bits. The high order M bits are encoded with the number of ALLO commands sent; the next *N-M* bits are encoded with the number of NOOP commands sent. According to formula 5, average $\frac{1}{2^{N-M}}\sum_{i=0}^{2^{N-M}-1} i$ NOOP commands are used to encode the low order *N-M* bits and average $\frac{1}{2^M}\sum_{k=0}^{2^M-1} k$ ALLO commands are used to encode the high order *M* bits. In addition, an ABOR is needed to represent the beginning of every *N*-bit covert information. So, the number of commands needed to transmit *N* bits covert information is:

$$\frac{1}{2^M}\sum_{k=0}^{2^M-1}\left(k+\frac{1}{2^{N-M}}\sum_{i=0}^{2^{N-M}-1} i+1\right) \tag{7}$$

The formula 7 can be simplified as the following:

$$\begin{aligned}
&\frac{1}{2^M}\sum_{k=0}^{2^M-1} k+\frac{1}{2^N}\sum_{k=0}^{2^M-1}\sum_{i=0}^{2^{N-M}-1} i+1\\
&=\frac{1}{2^M}\sum_{k=0}^{2^M-1} k+\frac{1}{2^{N-M}}\sum_{i=0}^{2^{N-M}-1} i+1\\
&=\frac{2^M-1}{2}+\frac{2^{N-M}-1}{2}+1\\
&=\frac{2^M+2^{N-M}}{2}
\end{aligned} \tag{8}$$

The formula 6 can be simplified as the following:

$$\frac{1}{2^N}\sum_{i=0}^{2^N-1}(i+1)=\frac{2^N+1}{2} \tag{9}$$

When $2^N >> 1$, formula 8 approximately equals to 2^{N-1}. (10)

Define the optimal encoding function $f(M)=\frac{2^{M-1}+2^{N-M-1}}{2^{N-1}}=2^{M-N}+2^{-M}$, where *M*<*N*. To compute *M* to make the value of $f(M)$ minimal, the *M* must satisfy the following relationship: $f'(M)=0$. $2^{M-N}\ln 2-2^{-M}\ln 2=0$. Resolve the above equation and get $M=\frac{N}{2}$. It means to send minimal commands, *M* must equal to *N*/2. The table below shows the covert channel encoding performance when *N*=8. The performance is the best when M=4.

Due to utilize the idle prevention scheme in FTP, the covert channel has good concealment. On the average, 16 commands are needed to transmit 8-bit covert information in best conditions. The bandwidth of the covert channel is 0.5 bit per command. The robustness is strong due to the retransmission scheme of TCP.

Table 1. The covert channel encoding performance when *N*=8 and *M* is selected from 0 to 8.

M	Average command number of covert channel 1	Average command number of covert channel 2
0	128.5	128.5
1	65	128.5
2	34	128.5
3	20	128.5
4	16	128.5
5	20	128.5
6	34	128.5
7	65	128.5
8	128.5	128.5

Note: The covert channel 1 is based on hierarchical sequence length and covert channel 2 is based on sequence length.

6 Conclusions

Two kinds of novel covert channels are studied in the paper. In fact, the analysis method about the FTP covert channels can be also used in finding covert channels in other application protocols. And we can also embed the authentication information into the application protocols when authentication is necessary. The implicit authentication is useful because it will not arouse the attacker's attention. The TCP/IP application layer covert channel is a new research area and there are still much more works need to be done.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Project Number: 60372052), the Foundation for the Author of National Excellent Doctoral Dissertation of China(Project Number: FANEDD-200238), the Multidiscipline Scientific Research Foundation of Harbin Institute of Technology (Project Number: HIT.MD-2002.11), the Foundation for the Excellent Youth of Heilongjiang Province, and the Program for New Century Excellent Talents in University.

References

1. J. Postel, J. Reynolds. File Transfer Protocol (FTP)[S]. rfc 959, 1985
2. Lu Dahang. The Research and Implementation of Covert Channel Based on Network Protocols[J]. Computer Engineering and Application, 2003,vol.02:183-6.(In Chinese)
3. Bruce Schenier, Applied Cryptography, protocols, algorithms, and source code in C, China Machine Press[M], 2000, p11-3. (In Chinese)
4. Giffin, J., R. Greenstadt, et al. Covert messaging through TCP timestamps[J]. Privacy Enhancing Technologies, 2003, 2482: 194-208.
5. Singh, A., O. Nordstrom, et al. Malicious ICMP tunneling: Defense against the vulnerability. Information Security and Privacy[C], Proceedings 2003, 2727: 226-236.

Data Hiding in a Hologram by Modified Digital Halftoning Techniques

Hsi-Chun Wang and Wei-Chiang Wang

Department of Graphic Arts and Communications
College of Technology, National Taiwan Normal University
hsiwang@cc.ntnu.edu.tw

Abstract. The objective of this research is to design a novel data hiding method for the dot-matrix hologram. The modified error diffusion techniques for addressing the halftone dots in six colors are proposed to make efficient use of the image area and to hide information. After outputting the encoded hologram, which is in analog format, an independent data recovery system is applied to capture the encoded holographic image and to extract the hidden information. The results show that it is feasible to hide data in a dot-matrix hologram. The method proposed in this research can achieve multiple security features for hologram in anti-counterfeiting applications.

1 Introduction

With the innovation of computer technology, copyright concerns in digital media contents have attracted numerous researchers in studying watermarking or information hiding techniques [1]. However, information hiding in binary or halftone images, which are often in printed (or analog and physical) format instead of electronic format such as digital image and audio, is less addressed and considered challenging. There are three difficulties regarding data hiding in image of physical format: binary image is lack of capacity for hiding data [6,7,8], the physical and chemical interaction in the output procedure is complex, and an independent procedure of data recovery is needed to extract the hidden data from the analog image. However, information hiding in printed binary image is important in designing and preparing security documents [2,8]. In recent years, high resolution desktop scanners/printers and color copiers become much more accessible for general public. On the other hand, the unauthorized duplication of security documents, such as ID, certificates and banknotes, is a crucial issue in many countries. Efforts have been devoted to the advances of anti-counterfieting technologies nowadays. Holography is one of the most effective ways to withstand illegal duplication because it can provide unique iridescent features which cannot be duplicated by existing digital copy techniques.

Holography was first found by Dennis Gabor in 1948 while the wave-front reconstruction was proposed [9]. The amplitude and phase of the interfered waves were recorded for later reconstruction. For holography, the light source should be coherent to achieve better performance. The invention of laser in 1960 did provide a good tool to boost holography technologies. For anti-counterfeiting application, the dot-matrix hologram, or embossed hologram, is widely used. Two light sources are focused on a photoresist, and the interference grating is recorded. The grating spot, composed of

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1086–1092, 2005.

interference fringes, is the smallest unit in a dot matrix hologram. The parameters of grating spot include spot size, grating orientation and grating pitch [10].

The organization of this paper is presented in following sections. In section 2, the fundamentals of halftoning technique are briefly reviewed. The modified error diffusion for data hiding in hologram is presented in section 3. The hidden data recovery system and some close-up images are shown in section 4. In section 5, the conclusions are presented.

2 Digital Halftoning

Halftoning has been an traditional and important printing process. A con-tone image has to be halftoned as a bi-level image because of the limited tones which output device can reproduce. Human eyes tend to smooth and integrate the bi-level image, and a continuous-tone image is reconstructed by the visual system. Since the advances of information technolgy, digital halftoning has been an active research area and it has been widely adopted by digital output devices [3,4,5]. In general, digital haftoning can be divided into two major categories: ordered dithering and error diffusion. The grayscale values, G(i,j), is converted into blackness values, b(i,j), in Eq. (1). Ordered dithering is performed by an 8x8 (or other size) thresholding matrix, T, on the blackness to render a halftone image y(i,j) in Eq. (2).

$$b(i,j)=1-G(i,j)/255 \tag{1}$$

$$y(i,j) = \begin{cases} 1, & b(i,j) \geq T \\ 0, & b(i,j) < T \end{cases} \tag{2}$$

Error diffusion is a well-known algorithm for frequency modulation, and it is first proposed by Floyd and Steinberg [3]. The kernel to diffuse the error is listed in Table 1.

Table 1. Floyd and Steinberg error diffusion kernel

	1	-7/16
-3/16	-5/16	-1/16

In Fig. 1 and Fig. 2, the typical halftone images rendered by ordered dithering and error diffusion are illustrated. The distance between halftone dots in Fig. 1 is fixed and the sizes of the dots are varied. They are also known as AM (amplitude modulation) halftone dots. For the image in Fig. 2, which is rendered by error diffusion, it shows the same size halftone dots, while the distance between halftone dots are varied. This is a typical example of FM (frequency modulation).

Comparing with Figs 1 and 2, it is obvious that FM halftone dots can provide better description of the image details. However, AM halftone dots still have advantages of less computational complexity and less dot gain [11]. As for color printing applications, the halftoning process is performed by CMYK channels, respectively and independently. The workflow is valid for printing but it cannot be adopted by full color hologram. This will be discussed in the following section.

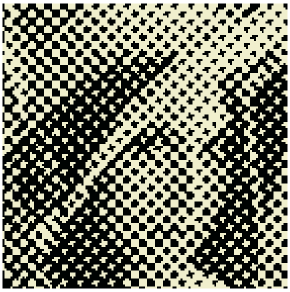


Fig. 1. Ordered Dithering

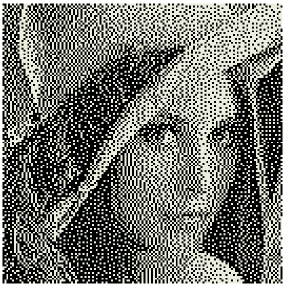


Fig. 2. Error Diffusion

3 Modified Halftoning Techniques for an Encoded Hologram

In this section, the modified halftoning technique exclusively for dot matrix color hologram is proposed. In dot matrix hologram, each pixel of the image area can be addressed by only one spot of red, green, blue, cyan, magenta, yellow, or none (as black), and the addressed color dots would not overlap. However, for halftone dots in color printing, dots of different colors are output separately and they do overlap.

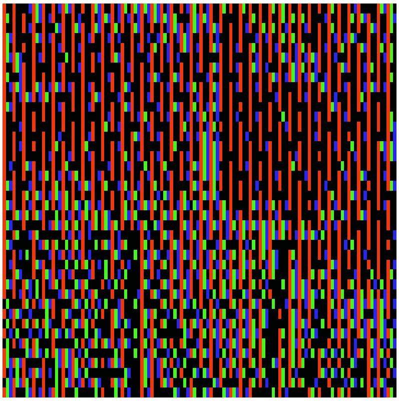


Fig. 3. Color halftone image in stripe geometry

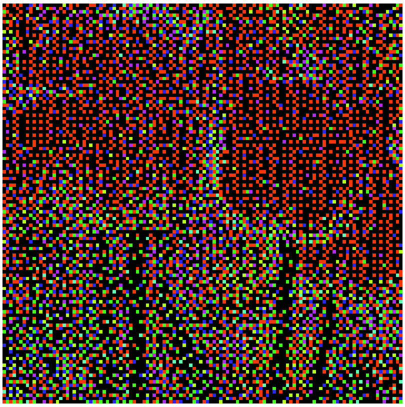


Fig. 4. Color halftone image by modified error diffusion.

In Fig. 3, the color halftone image in "stripe" geometry is illustrated. It is a common geometry for hologram to display color images and this geometry is similar to the way LCD displays images. The sub-pixel of R, G and B are addressed independently, but there is no gray-level for each channel. If all the three sub-pixels are turned on, it perceives like white light. Relationships of RGB and CMY are formulated in Eq. (3).

$$\begin{cases} G + B = C \\ R + B = M \\ R + G = Y \end{cases} \tag{3}$$

A novel algorithm is proposed to address R, G, B, C, M or Y color dot on a selected pixel, and the error is diffused to the neighboring pixels. The result is shown in Fig. 4 and it can provide much better descriptions of the details. It is a color and non-overlapped version of error diffusion algorithm which is especially useful for the output of dot-matrix hologram.

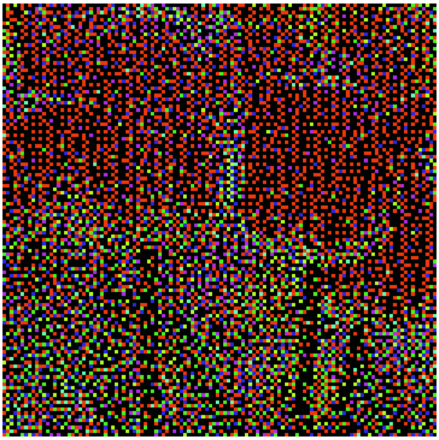


Fig. 5. Color halftone image with hidden data

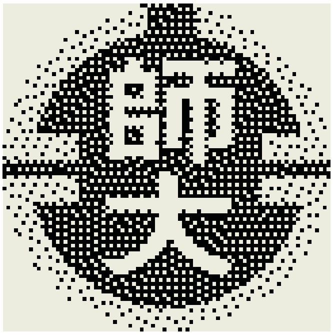


Fig. 6. Encoded watermark

Furthermore, our color error diffusion algorithm can be modified and incorporated with “data hiding error diffusion” technique proposed by Fu and Au [6]. We avoids the clustered 0’s and 1’s (lead to visual artifacts) generated from pseudo random number by isolating the encoded locations. The resulting encoded color halftone image is shown in Fig. 5. Actually, the full size of this color halftone image is 211x304, and the cropped image of 120x120 is shown in Fig. 5 for close-up observation. The encoded information, a logo pattern of NTNU sized 86x86 shown in Fig. 6, is hidden in the full size of the color halftone image in Fig. 5. During the encoding process, the pixel is forced to be black (0) and error diffuses to the neighboring pixels, while the encoded data is 0. The pixel is switched to R, G, B, C, M, or Y depending on which one has the stronger intensity, while the encoded data is 1. Comparing Figs. 4 and 5, it is very hard to tell the differences between them.

4 Hidden Data Recovery System

Since the encoded image on a hologram is in analog, instead of digital, format, it is necessary to design an independent data recovery scheme to capture the image on hologram and to recognize encoded data. It is subtle that the quality of image itself on the hologram or the substrate plays an important role during data recovery. That is, no matter how good the imaging apparatus and the recognition software are, if the image is not well-addressed on the substrates, the recognition results would not be satisfactory. In Figs. 7 and 8, we can compare the analog image quality on a hologram and that on a printed image. It shows that the dot matrix hologram in Fig.7 does provide good image quality and is almost free from “dot gain”[11] which is obvious in printed image in Fig. 8. Fig. 9 is an output dot matrix hologram of the full size watermarked image (211x304) shown in Fig. 5. The extracted watermark is shown in Fig. 10.

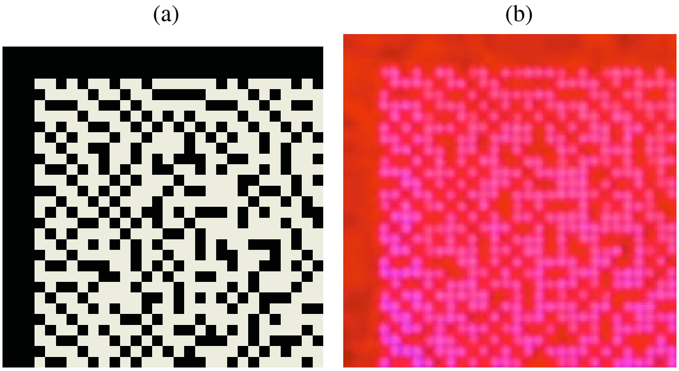


Fig. 7. (a) the original digital binary image. (b) the microscopic picture of a hologram output at 400dpi

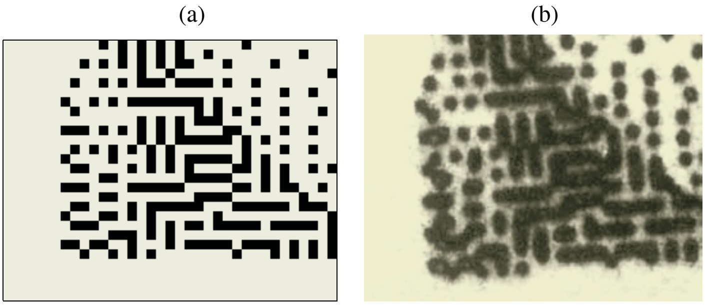


Fig. 8. (a) the original digital binary image. (b) the microscopic picture of a printed image output at 300dpi [8]

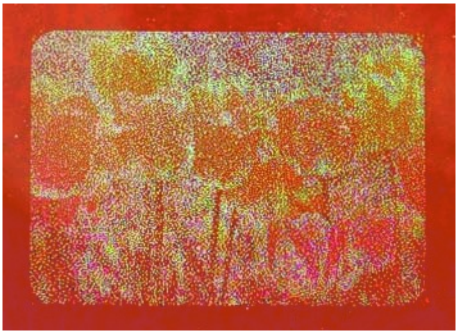


Fig. 9. An output dot matrix hologram

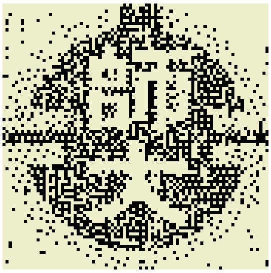


Fig. 10. Extracted watermark

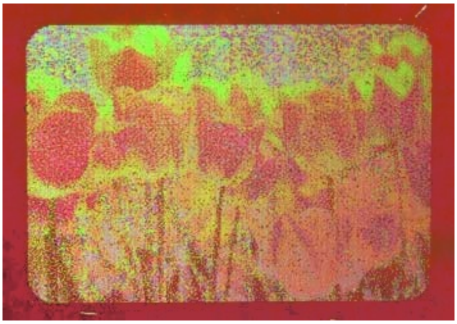


Fig. 11. A tone-adjusted image on a hologram

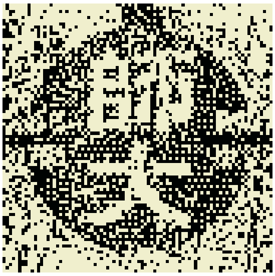


Fig. 12. Extracted watermark

To extract the watermark in an encoded hologram shown in Fig. 9 is a challenging task during data recovery. A high resolution digital camera with proper lighting condition is necessary to capture a good digital image. The geometrical transformation between the full 211x304 digital binary image and the captured image is constructed by the following equations (4)-(6)[12].

$$\begin{bmatrix} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 \\ v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \end{bmatrix} = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & a_4 & a_5 \\ b_0 & b_1 & b_2 & b_3 & b_4 & b_5 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ y_1 & y_2 & y_3 & y_4 & y_5 & y_6 \\ x_1^2 & x_2^2 & x_3^2 & x_4^2 & x_5^2 & x_6^2 \\ x_1y_1 & x_2y_2 & x_3y_3 & x_4y_4 & x_5y_5 & x_6y_6 \\ y_1^2 & y_2^2 & y_3^2 & y_4^2 & y_5^2 & y_6^2 \end{bmatrix} \tag{4}$$

$$\boldsymbol{U=AX} \tag{5}$$

$$\boldsymbol{A=UX^T(XX^T)^{-1}} \tag{6}$$

where (x_i,y_i) is the coordinates of the digital image ($\boldsymbol{X}$), and (u_i,v_i) is the coordinates of the captured image ($\boldsymbol{U}$) of the encoded hologram. $\boldsymbol{A}$ is the 2nd order transformation matrix which can be calculated first by selecting at least six sets of corresponding points. Then the corresponding grayscale values of each pixel in the digital image can be extracted, and a threshold is chosen to complete the recognition. A tone-adjusted image, which is addressed by much more grating spots, on a hologram is shown in Fig. 11, and a satisfactory watermark can be extracted in Fig. 12.

5 Conclusions

In this paper, a modified error diffusion technique was applied to information hiding especially for a hologram. The independent imaging and recognition system is capable of recovering the hidden data in the hologram. Hologram with this unique feature of data hiding can enhance the security and anti-counterfeiting functions. The techniques we proposed in this work have many value-added applications to the security printing. In our future works, the data compression and error correction code will be used to improve the data encoding capacity, as well as to withstand the unexpected attacks, such as particles, scratches, etc.

Acknowledgments

This research is supported by the National Science Council of the Republic of China under the grant of NSC93-2516-S-003-011. Thanks to AHEAD Optoelectronics, Inc. for technical assistances.

References

1. W. Bender *et al.*, Techniques for data hiding, IBM System Journal, vol 35, Nos 3&4, pp 313-336, 1996.
2. R.L. van Renesse, Optical Document Security, 2nd Ed., Artech House, London, 1997.
3. R.W. Floyd and L. Steinberg, An adaptive algorithm for spatial greyscale, proceedings of Society for Information Display, Vol 17(2), pp75-77, 1976.
4. R. Ulichney, Digital Halftoning, MIT Press, Cambridge, MA, 1987.
5. T.N. Pappas, J.P. Allebach, D.L. Neuhoff, "Model-Based Digital Halftoning", IEEE Signal Processing Magazine, pp.14-27, July 2003.
6. M.S. Fu and O.C. Au, “Data Hiding Watermarking for Halftone Images,” IEEE Transactions on Image Processing, 11(4)477-484, 2002.
7. M.S. Fu, O.C. Au, “Correlation-based watermarking for halftone images”, IEEE International symposium on circuits and systems conference, Vancouver Canada, May 2004.
8. H.C. Wang, An Information Hiding Technique for Binary Images in Consideration of Printing Process, PCM 2002, Lecture Notes in Computer Science 2532, pp. 460-467, 2002.
9. R.J. Collier, C.B. Burckhardt, L.H. Lin, Optical Holography, Academic Press, New York 1971.
10. A. Rhody, F. Ross, “Holography Market Place”, 8th Ed, Ross Books, 1999.
11. Z.A. Prust, Graphic Communications – the Printed Image, Goodheart-Willcox, Tinley Park, IL, 1989.
12. W.K. Pratt, Digital image processing, 2nd Ed, John Wiley & Sons, Inc., 1991.

A Secure Steganographic Scheme in Binary Image

Yunbiao Guo[1], Daimao Lin[2], Xiamu Niu[3], Lan Hu[2], and Linna Zhou[2]

[1] School of Electronic Information Engineering, Tianjin University, 300072, China
gybgnx@hotmail.com
[2] Beijing Institute of Electronic Technology Application, Beijing 100091,China
{hulan,wdmzln}@263.com
[3] Harbin Institute of Technology, Harbin 150001,China
xiamu.niu@hit.edu.cn

Abstract. A secure steganographic paradigm of binary image, which is fulfilled by simulating normal scanning process, is suggested in the paper. The indeterministic parts of scanning can be calculated by checking the differences between repeated scans. The information hiding is realized with reference to this indeterminism. We also provide a novel scheme to put this secure paradigm into practice by using the error control embedding technique. A more efficient practical method is presented based on morphological theory and the Matrix encoding strategy is utilized to improve the embedding performance.

1 Introduction

Steganography, derived from the Greek “covered writing”, is an emerging science of invisible communication. A steganographic scheme aims to hide the very existence of secret message by embedding them into innocuously looking cover data. The most relevant feature of steganographic scheme is the security, which refers to the difficulty that an adversary has in obtaining evidence or even grounds for suspicion of a confidential communication. The definition of perfectly secure stegosystem based on information theory can be found in [1].

A more challenging problem is to hide data in a two-color digital image because arbitrarily changing pixels in a binary image causes extremely noticeable artifacts. Compared with the plurality of proposed methods in steganography for the full color picture and video streams, the steganographic methods of the two-color image are very limited. Over the last few years, we have seen a growing number of papers proposing new techniques and ideas for binary image information hiding. However, not too much attention has been paid on the security of the previously proposed techniques. Some methods even do not provide any measurement to ensure good visual quality in stegocover [2]. A secure paradigm and a novel method for binary image steganography are presented in this paper. This new proposed method is to simulate a usual scan process while embedding. Elke Franze and Andreas Pfitzmann proposed a similar stego paradigm in [3]. They discussed stego system simulating a usual process, just like we will do, but do not go further in detail.

2 Embedding Scheme

As shown in fig. 1 the set of cover is **C**, the set of stego is **S,** the set of embedded message is **E** and the embedding processing is $\mathbf{f_E}$. If the mutual information

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1093–1100, 2005.

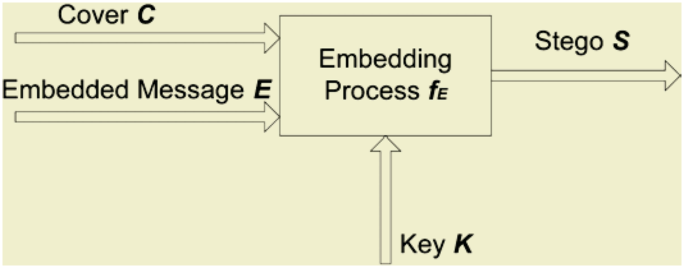


Fig. 1. The embedding model

$I(E,S|C)=0$, the stego system is secure system at the condition of cover-stego-attack, which means an attacker obtains no information whatsoever about the secret message by examining stego and cover.

Because $$I(E,S|C)=H(E|C)-H(E|(S,C)=0 \quad (1)$$

Then $$H(E|C)=H(E|(S,C)) \quad (2)$$

E and C can be thought as independent

$$H(E)=H(E|(S,C)) \quad (3)$$

So the condition of secure stego system is $H(E)=H(E|(S,C)$. Normally when an attacker knows the S and C, he can get the embedded message information by comparing S and C, $H(E)\geq H(E|(S,C)$. So as shown in [10], the secure problem cannot be achieved and the indeterminism of cover is introduced to fulfill secure embedding system.

However there may be an exception. If *C*=*S*,

Then $$H(C|S)=H(S|C)=0 \quad (4)$$

$$H(E|C)=H(E|S)=H(E) \quad (5)$$

It follows that $H(E)=H(E|(S,C))$. The secure condition can be met.

What does *C*=*S* mean? For the individual cover $c_i \in C$ and stego $s_i \in S$, it means $c_i = s_i$. That is impossible because c_i and s_i must have some differences to hold the embedded message. Nevertheless maybe we can get i and j, let the $c_i = s_j$ be achievable. In this specific situation, the *C*=*S* is a tenable necessary and sufficient condition for secure steganography.

The formula *C*=*S* means that the attacker obtains no information about whether the stego s_j contains the embedded information by observing *stego S* and *cover C* . Assume that C is infinite set, if $\forall e \in E, \quad \exists c_i \in C, c_j \in C \quad and \quad f_E(e,c_i)=c_j$, the secure condition *C*=*S* is achievable and f_E is the secure stego system.

The secure stego system f_E can be fulfilled by simulating the usual process that modifies data. The usual process must produce some differences between the input and the output, which can be used to embed the secret message.

The simulated process here is scan process. When we scan a text document, the repeated scanning of one image will result in several digital images that look very similar to the human eyes. However even if the same setting were used in scanning, the single pixel of images is not identical. That is the indeterminism of scans process, which ensure the security of the stego system, because the attacker is not able to decide the differences between the intercepted data caused by the embedding process or the simulated process.

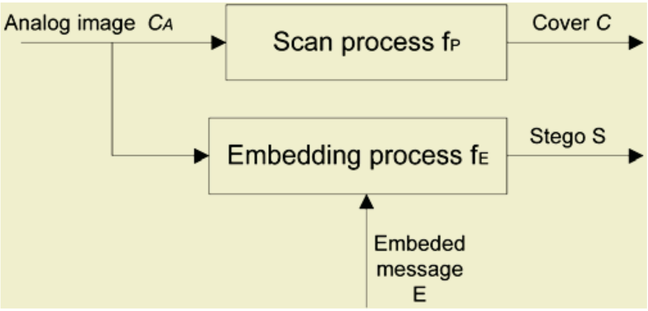


Fig. 2. Scan process and embedding process

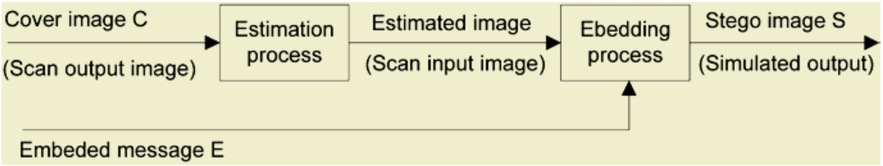


Fig. 3. The scan simulated embedding process

Fig. 2 shows the scanning and embedding process. The stego system is to simulate the scan process. If the simulated embedding function is elegant, the differences between the input image C_A and stego image S will match with the differences between the C_A and the cover image C(we call the scan output image cover). The attacker can not decide whether the differences are caused by the repeated scanning or the data embedding. C and S can be thought as equal. Normally the input image is analog and the stego system cannot directly access analog signal, hence the embedding process should be applied to the scan output image. Therefore before embedding we must do the input image estimating first. Fig. 3 show the scan simulated embedding diagram.

3 Scan Input Image Estimation and Error Control Embedding Method

Assume that a binary image X is the original cover, a number of outputs Xi (i=1…. l) can be obtained by repeated scanning X. It is obvious that Xi varies from each other because of the metrical noise.

Definition 1: Image distance, which means the differences between two images, is the sum of all differences between the corresponding pixels of two given images.

$$D = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \left| x_1(i, j) - x_2(i, j) \right| \tag{6}$$

Where the M and N are size of image and $x_1(i, j)$ and $x_2(i, j)$ are the corresponding pixels of two images.

Definition 2: Mean image is the average of all image, the pixel in mean image is

$$\overline{x}(i, j) = \frac{1}{L} \sum_{k=1}^{L} x_k(i, j) \tag{7}$$

Where the $x_k(i, j)$ is the pixel in number *k* image.

Here are the four steps to do embedding.

1. Scan the given image X with the same setting L times. We can get L output images X_1, X_2…X_L.
2. The mean image $\overline{X}$ can be easily worked out according to formula (7).
3. Calculate the distance between the mean image $\overline{X}$ and each output image using formula (6). We can get L Distances D_1,D_2,…D_L. The indeterminism of scan process can be described by the average of the distance $\overline{D} = \frac{1}{L} \sum_{k=1}^{L} D_k$.
4. Embed the secret data into mean image. Ensure that the noise introduced by the stegosystem keep pace with the scanning indeterminism $\overline{D}$.

Step 4 is the key part of the embedding scheme and it can be carried out using fixed partition method [4]. The binary image is partitioned into fixed blocks of size m by n; only one pixel can be modified to hold the secret message in each block. Hence the introduced error can be controlled by adjusting the block size to keep the embedding artifact and the scanning indeterminism at the same level.

We calculate the reciprocal of the average distance, $R = \frac{1}{\overline{D}}$. Let integer $K = 2^{\log_2 \lfloor R \rfloor}$ and K is easy to be decomposed into **m** by **n**. The image is divided into block of size m by n, embed the message in each block with matrix encoding [5].

Matrix encoding is a strategy that produces least conspicuous change to hold more secret bits. We can give a definition that holding efficiency is the ratio of the holding bits to the changing bits. Matrix encoding can increase the holding efficiency and decrease the changing rate much more. The following example can show that.

Assume that we use, a_1, a_2, a_3 three bits to hold two message bits x_1, x_2, Let $x_1' = a_1 \oplus a_3$, $x_2' = a_2 \oplus a_3$, changing not more than one bit can bring two according the following rule:

If $x_1 = x_1'$ and $x_2 = x_2'$, without any change;

If $x_1 \neq x_1'$ and $x_2 = x_2'$, modify a_1;

If $x_1 = x_1^{'}$ and $x_2 \neq x_2^{'}$, modify a_2;

If $x_1 \neq x_1^{'}$ and $x_2 \neq x_2^{'}$, modify a_3.

This example shows that at most changing one bit can carry 2 bits data. This example can be marked as (1,3,2). According to Hamming error-correcting method, the matrix encoding $(1,k,2^k - 1)$ can be easily got and its holding efficiency is $w = \frac{2^k}{2^{k-1} - 1} k$.

The block size should be big enough to ensure that embedding error is at the proper level. The entire pixel in the block can be used to carry the secret message; any one bit in the given block can be modified. This provides elegant condition for applying the matrix encoding method to increase the embedding payload.

It is necessary to point out that the embedding process is applied to the mean image. The mean image can be thought as the scan input image C_A or the estimation of the C_A; the stego system is to simulate the scan process so that the stego image looks like the scanning output. The secure stego system has been achieved.

4 Rapid Realization Based on the Mathematic Morphology Filter Method

From the proposed embedding scheme in the previous sector, we can see that the embedding process is complicated. The cover needs to scan many times and lots of calculations are needed to get the mean image and average distance of the scanned image. Furthermore, there may be some difficulty in the secret message extracting process. In order to simplify the embedding process, we try to estimate the mean image from single scanned image.

The main part of the noise introduced by scanning can be thought as metrical noise and the mean image of repeated scanning outputs can be seen as the original one. So the mean image estimation becomes a noise reduction problem. According to the experimental results, Visual quality of the scanning output image is not as good as that of the original one. Some edges, thin lines and small features are not very sharp and clear and the areas between these features are not smoothly varying. Salt noise and pepper noise are produced during scanning process. Mathematical morphology operation can eliminate this kind noise. We use mathematical morphology filter to do the mean image estimation.

An image is represented by a set of pixels. The morphological operations can be visualized as working with two images. The processed image is referred as the active image and the other kernel image is referred as the structuring element (SE). We can filter the active image by probing it with various SEs. The two basic major morphological operations erosion and dilation are defined as follows:

$$A \ominus B = \{A + X : X \in B\} \quad (8)$$

$$A \oplus B = \{X : B + X \in A\} \quad (9)$$

Where $\ominus$ and $\oplus$ denote the erosion and the dilation operation. *A* is the active image and *B* is the SE. The combination of the erosion and dilation operation leads to more complex morphological operations.

$$A \circ B = (A \ominus B) \oplus B \tag{10}$$

$$A \cdot B = (A \oplus (-B)) \ominus .(-B) \tag{11}$$

Where the opening operation ($\circ$) is defined as an erosion operation followed by a dilation using the same SE, and closing operation ($\cdot$) denotes as a dilation operation followed by the erosion using the same SE.

According to mathematical morphology theory, mathematical morphology filter can reduce salt and pepper noise by using opening and closing operations. Here the closing operation does the work more efficiently with the flat SE. We have done some experiments with the several SEs, the best result is obtained by the 2 X 2 square SE.

Once the mean image estimation is got by mathematic morphology filtering, the arduous repeated scanning and complicated calculation can be skipped. The embedding scheme can directly apply to the single scanned image as follow steps.

1. Filter the binary image (called reference image) by using closing operation with flat SE. The estimated mean image (this image can be thought as the original image) can be obtained.
2. Calculate ***D*** the distance between the estimated image and the reference image.
3. Work out the block size ***K*** using the distance ***D***.
4. Partition the estimated image according to blocks of size K. The message is embedded in each block with matrix encoding.

It is important to imply that embedding operation is applied to the estimated mean image. In this way the noise of the stego image and the reference image will be at same level.

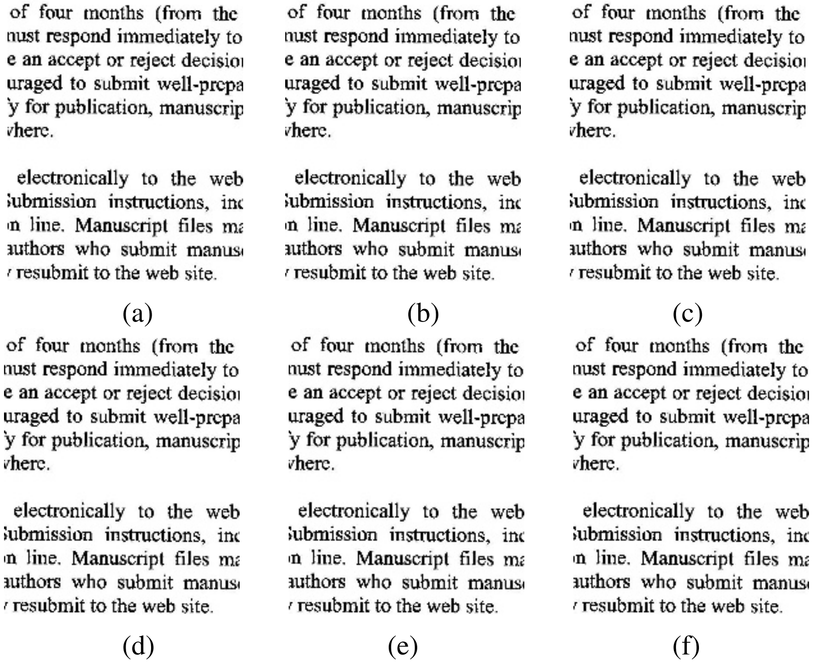
5 The Extract Procedure

Comparing with the embedding process the secret message extracting is much more simple. First it needs to filter the stego image with closing operation to get the estimated image (Although this estimated image from stego may have small difference with estimated image from reference image, these two image should have same noisy level) and it also needs to calculate ***D*** the distance between the estimated image and stego image. Then the block size can be worked out $K = 2^{\left\lfloor \log_2 \left(\frac{1}{D} \right) \right\rfloor}$. Once the bloke size ***K*** is obtained, the decode matrix can be decided and the secret message can be extracted with the matrix decoding.

Ensure that the embedding and extracting using identical block size and the encoding matrix and decoding matrix must be same. Then all block in the image should participate in the embedding process (If the secret message do not has enough length, made complement with random sequence) so that the stego image and reference image have same noisy level. Because the block size K is in round numbers of 2^n mode, even though the image distance ***D*** has little bitter error, the correct block size K could still be obtained in the most situations. In the real stego communication, the stego binary image do extraction test before sending to make sure that the secret message can be reliably transmited to the destination.

6 Experimental Results

The performance of the embedding scheme has been tested by simulation experimental results. Several outputs were obtained by scanning a single printing document image with same setting repeatedly. Filter the output image with mathematic morphology closing operation to estimate the mean image. Also difference ***D*** was worked out to control the block size. The embedding process was applied using double ***D***, ***D*** and half ***D*** to control the changing rate. Fig. 4 shows the visual effect of stego image.


of four months (from the
nust respond immediately to
e an accept or reject decisioı
uraged to submit well-prepa
y for publication, manuscrip
vhere.
electronically to the web
ubmission instructions, inc
n line. Manuscript files ma
uthors who submit manus
resubmit to the web site.

(a) (b) (c) (d) (e) (f)

(a) One typical image of 30 scan outputs,
(b) The mean image of the 30 scan outputs,
(c) The estimated image from (a) using morphology closing operation,
(d), (e), (f) Stego image form (c) using 0.5D, D and 2D to control the changing rate, D=0.00116, the block size K=2048,1024 and 512.

Fig. 4. Visual effect of the simulated embedding

We also did the simulation experiment with white image by repeatedly scanning a white paper. Not all pixels in the scanning output are white. A tiny part black pixels scatter in the whole image. Filter the outputs with morphology method the pure white image can be obtained. Conduct the embedding experiment as proposed scheme. The stego image looks like a pure white image as the scanning output. To some extend the white image can act as the stego cover as long as the embedding rate is suitable.

7 Conclusion

The embedding scheme simulating a usual process can construct perfect stegosystem. The difficulty may be how to fulfill this scheme. This paper presents a practical executive embedding model to implement the simulation. The present method ensures that the embedding artifact keeps pace with the scanning noise. There are some flaws in the embedding scheme such as disconsidering the distribution of the scanning noise. Even though the stegosystem simulates the scanning process only from the view of error, this opens up a direction in secure steganogrphy of binary images that needs to be further explored.

References

1. Thomas Mittelholzer, An information-Theoretic Approach to Stanganography and Watermarking, Lecture in computer science, Vol.1768 Springer-Verlag, Berlin Sep 2000 pp.1-16
2. Minya Chen, Edwar K.Wong, Nasir Memon and Scott Adams, Recent developments in document image watermarking and data hiding, Proceedings of SPIE Multimedia Systems and Application, Aug 2001, Vol.4518: pp.166-176.
3. Elke Franz and Andreas Pfitzmann, Steganography secure against cover-stego-attacks, Lecture in computer science, Vol.1768 Springer-Verlag, Berlin Sep 2000 pp.29-46
4. Yu-yuan Chen, Hsiang-kuang Pan, and Yu-Chee Tseng. A secure data-hiding scheme for two-color images, IEEE Transactions on communications, August 2002,Vol.50□8□pp:750-755.
5. Ron Crandall: Some notes on steganography. Posted on steganography mailing list, 1998 http://os.inf.tu-dresden.de/~westfeld/crandall.pdf
6. Yu-chee Tseng and Hsiang-kuang Pan. Data hiding in 2-color images, IEEE Transactions on computers, July 2002,Vol.51 (7): pp263-277.
7. K Tanaka and K. Matsui. Video-Steganography: How to Secretly Embed a Signature in a Picture, Proc. IMA Intellectual Property Project. 1994,Vol. 1 (1): pp187-205.
8. F.Cheng and A.N.Venetsanopoulos, “An adaptive morphological filter for image processing ” IEEE Trans. Image Processing, Oct 1992, vol 1, no 4, pp.533-539.
9. C.S.Regazzoni , A.N.Venetsanopoulos, G.L.Foresti, G.Vernazza, “Statistical Morphological Filters for Binary Image Processing,” Proc. Int Conf. Acoustics Speech and Signal Processing ICASSP ’94, Adelaide, Australia, 1994, pp.77-80.
10. J.Zöllner, H.Federath, H.Klimant, A.Pfitamann, R.Piotraschke, A.Westfeld, G.Wicke, G.Wolf, “Modeling the security of steganographic system”, Second international workshop Preproceedings, April 1998, pp15-17

A Reversible Information Hiding Scheme Based on Vector Quantization

Chin-Chen Chang[1,2] and Wen-Chuan Wu[2]

[1] Department of Information Engineering and Computer Science, Feng Chia
University, Taichung 40724, Taiwan, R.O.C.
ccc@cs.ccu.edu.tw
[2] Department of Computer Science and Information Engineering, National
Chung Cheng University, Chiayi 62145, Taiwan, R.O.C.
wenn@cs.ccu.edu.tw

Abstract. In this paper, we present a novel reversible information hiding scheme based on vector quantization (VQ) technique. First, VQ technique is used to encode the host image to form an index table. In the embedding phase, what the secret is hidden into the index table depends on the group id including the most closed codeword. And then, the modified index table is generated. In order to restore the original, we develop a flipping index table according to the original index table and the modified one. Finally, the stego-image could be constructed by combining the modified index table with the flipping one as well. Experimental results show that the performance of the proposed scheme is better than other VQ-based information hiding scheme in terms of the image quality and the distortion of the stego-image. Besides, in the proposed scheme, the original index table of the host image could be restored completely.

1 Introduction

Steganography [7] is known as information/data hiding technique, which secretly embeds data into a transmitted medium, called host media. And the medium that hidden secret, called stego-media, is transmitted to the receiver. When receiver gets the stego-media, he/she could pull the secrets out through the extraction algorithm. The stego-media in the steganography system must be similar to the original host media. For this reason, the stego-media would not be caught easily by intruders' attentions when transmitting on the network. Some possible attackers might be cheated by the camouflaged stego-media as well. Therefore, Steganography scheme is a better method to effectively deliver information or secret data without awareness.

In order to hide the secrets into the image, some nonessential contents in the host media are altered to camouflage. The conventional and simple way is LSB method, which is to slightly modify the least significant bits of pixel values. On the other hand, a great deal of works is to embed secret into the compressed code, such as VQ-compressed code [4,5,6,9], SOC-compressed code [2], and so on. In 2000, Lu et al. first proposed the concept of clustering the VQ codebook [6]. They divided the codebook into 32 groups, and each of them involves 16 similar codewords. In the embedding phase, the secrets are hidden by way of shifting to pick other codeword in the same group. However, this way has a serious defect that the method would need the

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1101–1107, 2005.

original image to extract secret data. Later, Lu et al. proposed the improved scheme [5] without the original image in the extraction procedure.

In 2002, Jo and Kim proposed a more effective and simpler data hiding scheme based on VQ-compressed code [4]. They took the two extremely similar codewords as a pair and threw them randomly into the different two groups. One is used to hide the secret bit 0, and the other is used to hide the secret bit 1. Finally, the codewords which could not make a pair are organized into the third group. This group represents no hiding secret data. For Jo and Kim's scheme, it is not only simple to cluster but also easy to extract information without the original image. Recently, numerous researches in this field aim at the reversible data hiding [1,3,8]. Reversibility indicates that the algorithm allows the embedding process to be reversed, that it restores the original image completely when the embedded data are removed. In this research, we refine Jo and Kim's scheme to be a reversible data hiding scheme. The rest is organized as follows. In Section 2, the previous VQ-based data hiding method are briefly reviewed. Then, in Section 3, the details of our proposed method are presented. In order to show the performance of the proposed method, some experimental results are given in Section 4. Finally, in the last section, some conclusions are made.

2 Related Works

In 2002, Jo and Kim proposed an information hiding scheme [4] based on vector quantization. This scheme modifies the VQ-index table to achieve the goal of data hiding. The image quality of modified index table is well resembled with the original one. In the beginning of running this scheme, the codebook C is divided into three sub-codebooks, SCB_0, SCB_1, and SCB_{-1}. The classification rule is that the two codewords, which are the most similar to each other, are treated as a pair, and the two paired codewords are placed randomly into the different groups, SCB_0 and SCB_1, respectively, by a key. Consequently, each codeword in one group, SCB_0 or SCB_1, can find its alternative codeword in the opposite group, SCB_1 or SCB_0. Here, the Euclidean distance formula is adopted for calculating the similarity degree. While the distance of two codewords is not greater than the specific threshold TH, it represents that these two are very similar. Finally, the remained codewords which could not be a pair would belong to other group, SCB_{-1}. The codewords in the group SCB_{-1} are used to represent that embedding no secret in the image block. Therefore, each codeword $C_i \in C$ would belong to one of the three sub-codebooks, SCB_0, SCB_1, and SCB_{-1}.

The following is the data embedding process of Jo and Kim's scheme:

Step 1: Host image is divided into non-overlapping blocks of $k \times k$, where $k = 4$.

Step 2: For each block, VQ encoding algorithm is processed to find out the most similar codeword C_i. And then obtain the index value i.

Step 3: Find out the group G_p including the codeword C_i. If p is equal to -1, it represents that the block could not embed any secret, and go to Step 2.

Step 4: If p is equal to b_t, then the index value i would be treated as the encoding code of the block, where b_t is one bit of the secrets. Otherwise, the index value j of the alternative codeword C_j of the codeword C_i is treated as the encoding code.

Step 5: Repeat Steps 2-4 above again and again until all the blocks have been processed to form an embedded index table T_s.

As for the secret extraction, first, the receiver has to classify the codewords in C into three sub-codebooks by using the classified key, which is only known between two parties. The extraction process of Jo and Kim's scheme can be stated as follows:

Step 1: Get out of one index value i from the compressed index table T_s.
Step 2: Find the codeword C_i, whose index is i, by using table-lookup technique. If C_i is in group SCB_{-1}, it represents that the block does not embed any secret. And go to Step 1.
Step 3: If C_i is in group SCB_0, it represents that the block embeds the secret '0'. On the contrary, the block embeds the secret '1' when C_i belongs to group SCB_1.
Step 4: Repeat Steps 1-3 above again and again until all the blocks have been taken to extract secrets.

3 The Proposed Method

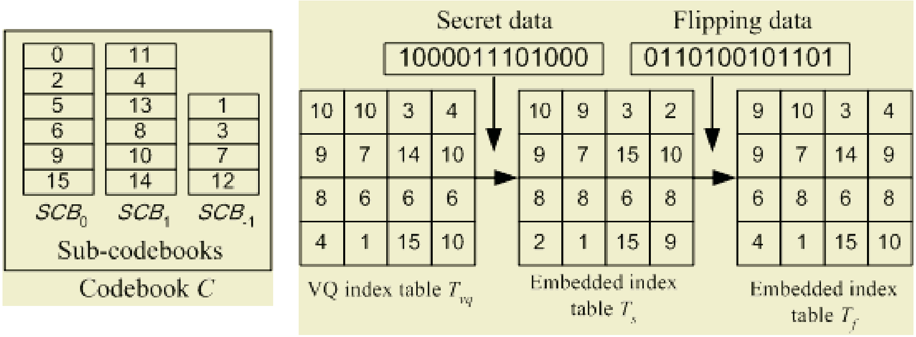
In this section, we will present the reversible data hiding scheme, which consists of the embedding and extraction procedures, based on VQ compression technique. The detailed description of these two procedures would be discussed in this section.

3.1 The Embedding Procedure

The goal of the embedding procedure is to embed the secrets into host image I and perform the reversibility. The proposed scheme is based on VQ technique by hiding secrets in the VQ compressed code. Hence, the first half of the embedding procedure adopts Jo and Kim's scheme, and the rest of it is to hide the reversible information, which is originally restored. First, host image I is encoded into an index table T_{vq} by using VQ technique. And then, embed the secrets into T_{vq} by using Jo and Kim's scheme to obtain a table, named as the hidden index table T_s. However, the table T_s could not restore the original T_{vq}. In order to achieve the reversion, we attempt to record the changes between T_{vq} and T_s to restore the original table T_{vq} in the future procedure of extraction. Hence, we treat the usage action of the alternative codeword in the secret embedding as the flipping information. We record the flipping bit as '1' while using the alternative codeword. On the contrary, we record the flipping bit as '0' while the values in T_{vq} and T_s are identical. As for the blocks embedded no secret, we would not record any information because the compressed code would not be changed. In the end, we would receive a sequence of flipping bits, as reversible information, which could be used to restore the original VQ index table T_{vq}.

The sequence of flipping bits that we just mentioned as above would be embedded into the hidden index table T_s by using the same embedding algorithm, which is Jo and Kim's secret embedding scheme, to generate another table, the flipped index table T_f. However, we discovered that the table T_f only extracts the embedded flipping bits, but the secret could not be obtained at this time. Consequently, we would organize the two index tables T_s and T_f to develop a stego-image because the hidden table T_s and the flipping table T_f own the secret and the reversible information, respectively. The stego-image embeds not only the secret data, but also the flipping-bits information. For the way of organization, we integrate the codewords in the table T_s with the table T_f to be a k^2-dimensional vector. That is to say, $\frac{C_i \times 1 + C_i^{'} \times 2}{3} = \left(\frac{c_{ij} \times 1 + c_{ij}^{'} \times 2}{3} \right)$, where c_{ij} represents j-th component of codeword C_i in table T_s, $c_{ij}^{'}$ represents j-th component of codeword $C_i^{'}$ in table T_f, and $1 \leqq j \leqq k^2$. In this case, the secrets could be extracted, and the flipping bits embedded in stego-image could be used to restore the original table.

Fig. 1 shows an example of the embedding procedure in the proposed scheme. Suppose that the codebook used in this example has already been divided into three groups, shown in Fig. 1(a). Then, host image is encoded into the VQ index table T_{vq}, shown in Fig. 1(b). Second, the secrets are embedded into T_{vq} by using Jo and Kim's scheme to form table T_s. From tables T_{vq} and T_s, we know that the blocks 2, 4, 7, 10, 12, 13, and 16 change their compressed codes in order to embed the secrets. Because the first block has not changed its code, we would give the flipping bit '0'. Yet the second block has changed it, the flipping bit is '1', and so on. According to the above information, we could derive a sequence of flipping data '0110100101101' as the restoration data, and the data is also embedded into table T_s. Finally, we would obtain another index table T_f, and tables T_s and T_f would be used to organize the stego-image.



(a) The classified codebook (b) The embedding process

Fig. 1. An example of the embedding procedure

3.2 The Extraction Procedure

The goal of the procedure on extraction is to pull out the secret data, which is embedded in the stego-image, and to restore the original VQ index table of the host image. The codebook C is shared by the sender and the receiver. For a start, the receiver would adopt the same way of classification in order to divide the codebook into three sub-codebooks, SCB_0, SCB_1, and SCB_{-1}. Paired codewords are randomly thrown to groups SCB_0 or SCB_1 by the same classified key. Because the key is known in advance by these two parties, receiver and sender, the receiver has to generate three sub-codebooks which are the same as those needed by the sender.

In the phase of extraction, the receiver partitions the obtained stego-image to non-overlapping blocks whose size is $k \times k$, where $k = 4$. Each block X uses the technique of VQ encoding to look for the most similar codeword C_i in the codebook C. If codeword C_i is a member of group SCB_{-1}, it represents that the block X does not embed any secrets. Or we would operate the paired codewords to extract the secret data. The codewords of each pair C_i and C_j, where codeword C_j is the alternative of codeword C_i, are integrated proportionally to check if the values are completely identical with the pixel values in block X. If it is, it represents that the block X is compressed by the paired codewords in the embedding phase.

The proportional integration has four different possible conditions in the embedding phase: (1) $\frac{C_i \times 1 + C_i \times 2}{3}$, (2) $\frac{C_j \times 1 + C_j \times 2}{3}$, (3) $\frac{C_i \times 1 + C_j \times 2}{3}$, and (4) $\frac{C_j \times 1 + C_i \times 2}{3}$. If the value computed by condition (1) is completely identical to the pixel values in block X, it means that the index values of the block X in tables T_s and T_f are both I; if the value computed by condition (2) is completely identical to the pixel values in block X, it means that the index values of the block X in tables T_s and T_f are both j; if the value computed by condition (3) is completely identical to the pixel values in block X, it means that the index values of the block X in tables T_s and T_f are i and j, respectively; if the value computed by condition (4) is completely identical to the pixel values in block X, it means that the index values of the block X in tables T_s and T_f are j and i, respectively. Consequently, we could individually retrieve the index values in tables T_s and T_f.

After tables T_s and T_f are exactly retrieved, the secret data could be further extracted, and the original index table T_{vq} of the host image could be recovered. Since the table T_s is the result of embedding secrets, we could derive the secrets from table T_s by using the algorithm of extraction in Jo and Kim's scheme. And the flipping information could be obtained from table T_f because table T_f is the result of embedding flipping bits. Next, we are able to get the content of the original VQ index table T_{vq} according to the flipped information and the table T_s.

4 Experimental Results

In our experiments, we picked out six 256×256 images with 8 bits/pixel resolution as host images to evaluate the performance of the proposed reversible hiding scheme, and compare it with Jo and Kim's scheme. In addition, the basic codebook of size 256 is training by using LBG algorithm. First, the codebook is divided into three groups, SCB_0, SCB_1, and SCB_{-1}, where the paired codewords in SCB_0 and SCB_1 are extremely similar to each other. Then, these host images are individually divided into 4096 blocks of 4×4 pixels for VQ encoding algorithm. In our proposed scheme, we randomly generate a series of secrets by a pseudo number generator as so to measure the maximum capacity. Next, after the embedding procedure of our proposed scheme, where TH=80, each host image will become the stego-image, which carries secrets.

In this article, we still adopt two other measure criterions to describe the performance of the proposed scheme, except the stego-image quality in PSNR value, which are the extra distortion introduced (ED) and the number of secret bits embedded (NB). ED is defined as follows, ED=$MSE(I,I_s)$–$MSE(I,I_{vq})$, where I represents the original host image, I_s and I_{vq} are the stego-image and the VQ-encoded image, respectively, and *MSE* function is the mean-square error of the two images. Generally speaking, it is better if the fewer extra distortion is introduced by the embedding procedure. Table 1 shows the comparison between the proposed scheme and Jo and Kim's scheme in terms of ED, NB, and the quality. Here, the threshold *TH* in Jo and Kim's scheme and the proposed scheme is 80 to classify the codewords to produce the three sub-codebooks. And the sizes of the three sub-codebooks SCB_0, SCB_1, and SCB_{-1} are 114, 114, and 28, respectively. As shown in Table 1, the total embedding size in our proposed scheme is the same as Jo and Kim's scheme. Somehow, the image quality of stego-image in our proposed scheme is better than Jo and Kim's. It is 0.4 dB better than the comparison in average. Therefore, the performance of the proposed scheme is better than Jo and Kim's scheme by the above objective measurement.

Table 1. The comparison between the proposed scheme and Jo and Kim's scheme (*TH* = 80)

Images	VQ	Jo and Kim's Scheme			Proposed Scheme		
	PSNR (dB)	ED	NB (bit)	PSNR (dB)	ED	NB (bit)	PSNR (dB)
Boat	22.018	86.15	3866	21.187	63.85	3866	21.387
Baboon	25.851	75.85	3803	24.241	52.79	3803	24.671
Jet	26.563	61.91	3785	25.005	43.96	3785	25.402
Lena	26.518	64.2	3969	24.926	47.34	3969	25.291
Pepper	27.027	69.21	3893	25.161	52.01	3893	25.555
Tiffany	27.739	90.36	4043	25.125	59.53	4043	25.853

Table 2. The image quality comparison among different thresholds *TH*'s

Images	VQ	Jo and Kim's Scheme				The proposed Scheme			
		TH=50	*TH*=80	*TH*=100	*TH*=150	*TH*=50	*TH*=80	*TH*=100	*TH*=150
	PSNR	PSNR	PSNR	PSNR	PSNR	PSNR	PSNR	PSNR	PSNR
Boat	22.018	21.669	21.187	21.000	20.987	21.722	21.387	21.244	21.227
Baboon	25.851	25.205	24.241	23.854	23.661	25.382	24.671	24.356	24.184
Jet	26.563	25.857	25.005	24.723	24.610	26.033	25.402	25.143	25.024
Lena	26.518	25.681	24.926	24.509	24.485	25.846	25.291	24.998	24.889
Pepper	27.027	26.120	25.161	24.870	24.550	26.274	25.555	25.315	25.079
Tiffany	27.739	26.749	25.125	24.937	25.681	26.864	25.853	25.663	26.171
Average NB (bits)		3358	3893	4036	4092	3358	3893	4036	4092

TH= 50: $|SCB_0|$= 81, $|SCB_1|$= 81, $|SCB_{-1}|$=94; TH= 80: $|SCB_0|$=114, $|SCB_1|$=114, $|SCB_{-1}|$=28; TH=100: $|SCB_0|$=124, $|SCB_1|$=124, $|SCB_{-1}|$= 8; TH=150: $|SCB_0|$=127, $|SCB_1|$=127, $|SCB_{-1}|$= 2;

The more detailed discussion on the threshold value *TH* is shown as follows. In Table 2, where |•| denotes the object size, the quality in Jo and Kim's scheme and the proposed scheme are both worse while the larger threshold value *TH* is set. However, the image quality of the proposed scheme is better than that of Jo and Kim's scheme, and it is also close to the VQ-encoded image quality. Nevertheless, how to set up the threshold value *TH*? We can give some advices as shown in Table 2. According to Table 2, we found that the number of embedded secret data is more when the larger threshold value *TH* is set. Nevertheless, if more number of embedded secret data is given, there would also be relatively more distortions on the image quality. In order to balance the trade-off, we suggest that the threshold value is *TH*=80. In this case, the quality of stego-image is better than 25dB, and the ratio of the number of embedded blocks to the total number of blocks is over 95% in the proposed scheme.

5 Conclusions

In Jo and Kim's scheme, they did offer a simple and effective scheme to embed secret data into VQ index table. In this paper, we extend their concept above and propose a novel VQ-based data hiding method to accomplish the way of reversibility. In other words, the original VQ index table of the sender could be restored by the extraction algorithm. Experimental results reveal that our proposed scheme is superior to Jo and Kim's scheme in terms of the quality and the distortion of stego-image in the same hiding size. In addition to this, our proposed scheme is not only simple and effective but also able to restore the original index table to achieve the result of reversibility.

References

1. Alattar, A. M.: Reversible Watermark Using the Difference Expansion of a Generalized Integer Transform, *IEEE Trans. on Image Processing*, Vol. 13, pp. 1147-1156 (2004).
2. Chang, C. C., Chen, G. M., and Lin, M. H.: Information Hiding Based on Search-Order Coding for VQ Indices, *Pattern Recognition Letters*, Vol. 25, pp. 1253-1261 (2004).
3. Celik, M., Sharma, G., Tekalp, A., and Saber, E.: Reversible Data Hiding, *Proc. of the IEEE Int. Conference on Image Processing*, Rochester, New York, Vol. 2, pp. 157-160 (2002).
4. Jo, M., and Kim, H.: A Digital Image Watermarking Scheme Based on Vector Quantization, *IEICE Trans. on Information and Systems*, Vol. E85-D, pp. 1054-1056 (2002).
5. Lu, Z. M., Pan, J. S., and Sun, S. H.: VQ-based Digital Image Watermarking Method, *Electronics Letters*, Vol. 36, pp. 1201-1202 (2000).
6. Lu, Z. M., and Sun, S. H.: Digital Image Watermarking Technique Based on Vector Quantization, *Electronics Letters*, Vol. 36, pp. 303-305 (2000).
7. Petitcolas, F. A. P., Anderson, R. J., and Kuhn, M. G.: Information Hiding—a Survey, *Proc. of the IEEE*, Vol. 87, pp. 1062-1078 (1999).
8. Tian, J.: Reversible Data Embedding Using a Difference Expansion, *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 13, pp. 890-896 (2003).
9. Yu, Y. H., Chang, C. C., and Hu, Y. C.: A Steganographic Method for Hiding Data in VQ Encoded Images, *Proc. of the Int. Symposium on Intelligent Multimedia, Video and Speech Processing*, Hong Kong, pp. 358-361 (2004).

Zero-Based Code Modulation Technique for Digital Video Fingerprinting

In Koo Kang[1], Hae-Yeoun Lee[1], Won-Young Yoo[2], and Heung-Kyu Lee[1]

[1] Department of EECS, Korea Advanced Institute of Science and Technology,
Guseong-dong, Yuseong-Gu, Deajeon, Korea (Republic of)
ikkang@mmc.kaist.ac.kr
[2] Digital Contents Research Division, Electronics and Telecommunications Research
Institute, 161 Gajeong-dong, Yuseong-gu, Daejeon, 305-700, Korea
zero2@etri.re.kr

Abstract. Digital fingerprinting is a technique to protect digital contents from illegal reproduction and redistribution by marking unique information for individual user. A powerful but simple attack to diminish fingerprint signals is averaging. While several fingerprinting schemes against collusion attacks were proposed, they often do not fully account for multimedia data. In this paper, we propose a zero-based code modulation method which fully reflects marking assumption concept to embedding and detection of signals and anti-collusion code working mechanism as well. By manipulating 0-bit information of binary codes, detection accuracy and code separation under averaging attacks were enhanced. To demonstrate our method, we used an averaging-resilient fingerprint code based on GD-PBIBD theory and applied it video fingerprinting systems. Through the experimental results, we convince our method improves averaging resiliency of fingerprinting systems.

1 Introduction

Digital fingerprinting is a copyright protection technique that traces illegal distributions of copyrighted contents. A unique fingerprint code that identifies a recipient is inserted into host contents. When copies of the content are found from illegal routes, the content seller can identify traitors who had distributed the content by extracting the embedded fingerprint code. This technique can be employed as a part system in DRM applications [1]. Averaging attack is a serious problem in digital fingerprinting applications. The averaging attack is an attempt to remove the inserted fingerprints by averaging several copies of the content.

There are two main categories for averaging attack-resistant fingerprinting techniques. One is orthogonal modulation fingerprinting techniques; each user is assigned a spread spectrum sequence which is mutually orthogonal to each other as a fingerprint. The other approach is known as coded modulation methods that use collusion-resistant fingerprint codes. Boneh et al. introduced a code working concept named "Frame-proof code" that prevents false alarm of recipients who did not join in averaging attacks up to *c* colluders [2]. Trappe presented a fingerprinting architecture that adopted the frame-proof idea using BIBD (Balanced Incomplete Block Design)

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1108–1114, 2005.

theory [3]. Most of coded modulation methods are working based on marking assumption which assuming colluders can only change fingerprint code bits in which they have different values [2]. Since most of existing coded modulation works focus on generic data, they do not explore special properties of multimedia data for fingerprint system design.

In this paper, we propose a zero-based code modulation method against the averaging attack, which fully reflects the marking assumption to multimedia data. Our method modulates only 0 bit information in fingerprint code so that signal interruptions and signal strength diminution were eliminated. Our experimental results show the proposed method can efficiently embed and detect the fingerprints under averaging attacks along with multimedia data.

This paper organized as follows. Section 2 addresses major features of orthogonal modulation methods and coded modulation methods to figure out their problems. In section 3, we describe our fingerprint code scheme and propose the zero-based code modulation method. Its effectiveness is demonstrated in Section 4 through experiments and conclusion is presented in Section 5.

2 Previous Fingerprinting Schemes

2.1 Orthogonal Modulation Method

Orthogonal modulation approach is a straightforward method for digital fingerprinting. In this method, ***N*** orthogonal signals are used to accommodate ***N*** users, i.e. a mutually orthogonal signal is assigned to each user as a fingerprint. To decide the content owner or to trace colluders who made illegal copies, the same number of correlation as the number of users is required so that the detection complexity is high.

An additional drawback is that when the number of contents involved in the averaging attack increases, the strength of orthogonal signals is attenuated deservedly so that the detector cannot correctly trace colluders. Assume that correlation is used as a detection statistics and ***M*** users join to the coalition to make illegal copy ***V***. The signal of each user is orthogonal so that correlation values will be decreased in inversely proportional to the number of colluders ***M*** as follows.

$$\begin{aligned} L.C. &= \frac{1}{N}\sum W[i]\cdot V[i]\,,\ \textit{where}\ V = \frac{1}{M}\sum_{M} C_j\ . \\ &= \frac{1}{NM}\sum W[i]\cdot C[i]_j \end{aligned} \tag{1}$$

This shows a limitation of detector when orthogonal signals are averaged and we can expect when the large number of users joins to make illegal copies, orthogonal modulation scheme will fail to retrieve colluders successfully. More detail study is well-defined in [4].

2.2 Coded Modulation Method

The coded modulation method was designed to accommodate more users than that of orthogonal method with the same amount of signals. Trappe proposed a representative coded modulation method [3]. The binary fingerprint codes derived from BIBD theory serve up to ***c***-user averaging resistance. Averaging of up to ***c*** codes results in a unique code which can identify all codes associated with the averaging. A fingerprint signal for one user is composed of code bits and orthogonal signals as follows.

$$w_j = \sum_{i=1}^{M} b_{ij} u_i \qquad (2)$$

where $\boldsymbol{b_{ij}}$ are {±1}, which is multiplied by $\boldsymbol{u_i}$ according to ***i-th*** bit of ***j-th*** user code and $\boldsymbol{u_i}$ are orthogonal signals for ***i-th*** bit position. At the detector, $\boldsymbol{b_{ij}}$ are determined by correlation between $\boldsymbol{w_j}$ and $\boldsymbol{u_i}$ with a threshold value. When several fingerprint codes are averaged, the detected code is a bit-wise logical AND of those fingerprint codes. By comparing bit positions where the value is 1 in averaged code with a code book, the detector can find out which codes are involved in the alliance.

One problem of this method is a difficulty to decide a threshold value. What the detector would concern is magnitudes of detected values. Let suppose 7 users join to the averaging; one code begins with (0,1,…) and all the others begin with (1,1,…), respectively. In order to trace colluders correctly, the colluded code should be recognized as (0,1,…) for the first two bits. However, extracted results could be different from the theoretical result in blind detection. As shown in Fig. 3, some correlation results for the first bit represent bigger values than that of second bit, and that means detector cannot decide the constant threshold value to separate two groups. Mess of correlation points results in incorrect binary result codes and detector will deservedly report innocent users as pirates. For this reason, the embedding method of equation 2 may not guarantee the marking assumption; the main working mechanism of averaging-resilient fingerprint code.

3 Zero-Based Code Modulation Technique

3.1 Anti-collusion Fingerprint Code and Embedding

Anti-collusion fingerprint codes have a resiliency property for averaging attacks. Averaging attack on anti-collusion codes assumes that the result code is bitwise AND operation of those codes [3]; if at least one 0 is included in the codes at same bit positions, the result code bit at that position should be 0 no matter how many 1s are included in other codes. We used $(v, b, r, k, \lambda_1, \lambda_2)$ GD-PBIBD (Group Divisible Partially Balanced Incomplete Block Design) theory to generate our test code for averaging attack-resiliency [5][6]. In $(v,b,r,k,\lambda_1,\lambda_2)$ GD-PBIBD codes, the ***k*** indicates the number of 0s in a ***v***-length code and also means the number of patterns to be embedded in our scheme. The ***k*** should be a smaller number than ***v*** [6] and that means our code modulation method is more efficient than that of conventional coded modulation method. For tests, we adopted (72,81,9,8,0,1) GD-PBIBD. The length of each fingerprint code is 72 bits and total 81 users can have a unique fingerprint code; each fingerprint code has bit 0s at different positions each other. This code scheme can trace up to 7 users when they deliver averaging attacks with their own codes.

In a proposed scheme, the embedder modulates only 0 bit information in codes to fingerprint signals and inserts them to a host data. By utilizing only 0 bit information in codes, the embedder handles only core factors of colluder tracing mechanism and the marking assumption. The way of ***M*** bits-length fingerprint code modulation is depicted in Fig. 1 and we embedded those signals to host data using commonly used spread spectrum embedding method [7].

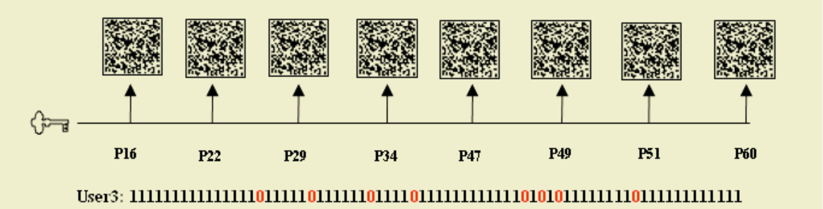

P16
P22
P29
P34
P47
P49
P51
P60
User3: 111111111111111011111011111101111011111111111101010111111110111111111111

Fig. 1. Zero-Based Code Modulation

3.2 Detection Process

The detection steps are composed of three stages; estimation of fingerprint signals, examination of fingerprint codes and tracing of pirates. At the estimation step, the detector extracts expected noise-like fingerprint signals from corrupted contents. Because noise-like orthogonal signals are added to original contents, some noise-pass filters are available. At the next step, our detector examines which indexes' patterns are embedded in the estimated pattern through a normalized correlation strategy [7] as depicted in Algorithm 1. Since our fingerprint code has fixed number of 0 bit information, ***k***, what the detector has to know is indexes of embedded signals. In case of the number of 0s are over the number ***k***, the content must have been attacked by the way of averaging because the codes cannot have ***k*** number of 0 bits at exactly same positions. At the final step, we can find out the pirates by examining the fingerprint code book and the result code. The colluder tracing method is the same way as that of conventional system [3].

```
1. Initialize result code bits $R_1, R_2, ......, R_M$ to 1
2. For estimated signals $E_1, E_2,...E_K$ and detector generated signals $S_1, S_2, ..., S_M$
   - calculate correlation between $E_i$ and $S_j$ $(1 \leq I \leq K, 1 \leq j \leq M)$
   - if correlation > threshold then set $R_j$ to 0 and resume correlation between $E_i$ and $S_{j+1}$
        - else if correlation < threshold then resume correlation between $E_i$ and $S_{j+1}$
```

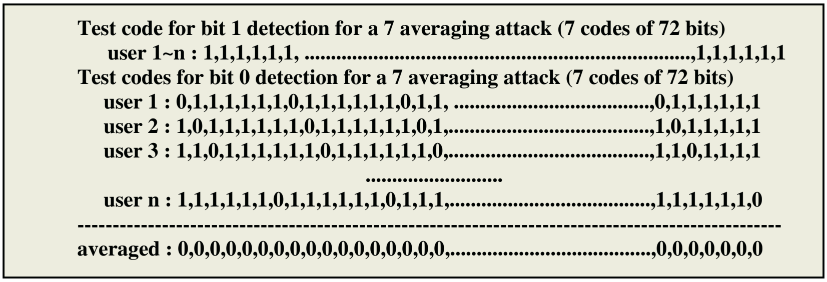
Algorithm 1. Fingerprint detection and code decoding

4 Experimental Performance

4.1 Code Separation Experiments

We will show experiments of detection value separations and results. This experiment is aimed to see how well the detection values are separated because the reason marking assumption does not work on multimedia is due to unclear separation of detected values. For this purpose, we made a fake code to demonstrate the worst cases of averaging attacks as shown in Fig 2. In a case of 0 bit detection tests, only one, among ***n*** bits at the same bit position in ***n*** codes, has a bit 0 and the others have 1. Theoretically, all bits of averaged code are 0.

We illustrated detection results of both systems in Fig 3. In (a), (c) and (e), we can see the detected values are getting blended up as the number of collusion increases. This situation happens because as the signals for bit 1are increased in equation 2, the linear correlation values are also increased. However in the proposed scheme, correlation points are clearly divided into two groups and the threshold can be easily determined. Even though detected values for bit 0 are decreased as collusions are increased, we can see two obvious partitions even in 7 collusion case. Therefore, our proposed scheme enhances code detection capabilities to decode the averaged code and to capture the colluders who delivered averaging attacks on multimedia data.



```
Test code for bit 1 detection for a 7 averaging attack (7 codes of 72 bits)
    user 1~n : 1,1,1,1,1,1, ...................................................................,1,1,1,1,1,1
Test codes for bit 0 detection for a 7 averaging attack (7 codes of 72 bits)
   user 1 : 0,1,1,1,1,1,1,0,1,1,1,1,1,1,0,1,1, ....................................,0,1,1,1,1,1,1
   user 2 : 1,0,1,1,1,1,1,1,0,1,1,1,1,1,1,0,1,.....................................,1,0,1,1,1,1,1
   user 3 : 1,1,0,1,1,1,1,1,1,0,1,1,1,1,1,1,0,.....................................,1,1,0,1,1,1,1
                                   ..........................
   user n : 1,1,1,1,1,1,0,1,1,1,1,1,1,0,1,1,1,.....................................,1,1,1,1,1,1,0
-------------------------------------------------------------------------------------------------
averaged : 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,.....................................,0,0,0,0,0,0,0
```

Fig. 2. Fake codes for 0 and 1 bit detection experiments

4.2 Colluder Tracing Experiments on Video

We applied our approach to video data extracted from DVDs and measured the performance of tracing colluders. We used fingerprint code from (72,81,9,8,0,1) GD-PBIBD and CIF (352x288) sized video data as host signals.

For the video fingerprinting performance assess, we averaged uniquely fingerprinted contents from 2 collusions to 7 collusion cases 100 times, respectively, and extracted the colluders. Table 1 shows the result of test; the number of correctly extracted colluders for each collusion case. When the number of colluders was below 4, we could trace all colluders successfully. When the colluder number exceeded 4, our detector partially failed to catch all colluders, but could trace at least one user and there was no report on innocent users.

Table 1. Detection results from 10 illegal copies of 10 videos for each collusion case

	Number of extracted colluders						
	1	2	3	4	5	6	7
1 collusion	100	-	-	-	-	-	-
2 collusion	-	100	-	-	-	-	-
3 collusion	-	-	100	-	-	-	-
4 collusion	-	-	-	100	-	-	-
5 collusion	-	-	-	3	97	-	-
6 collusion	-	-	-	1	3	94	-
7 collusion	-	-	1	1	2	6	90

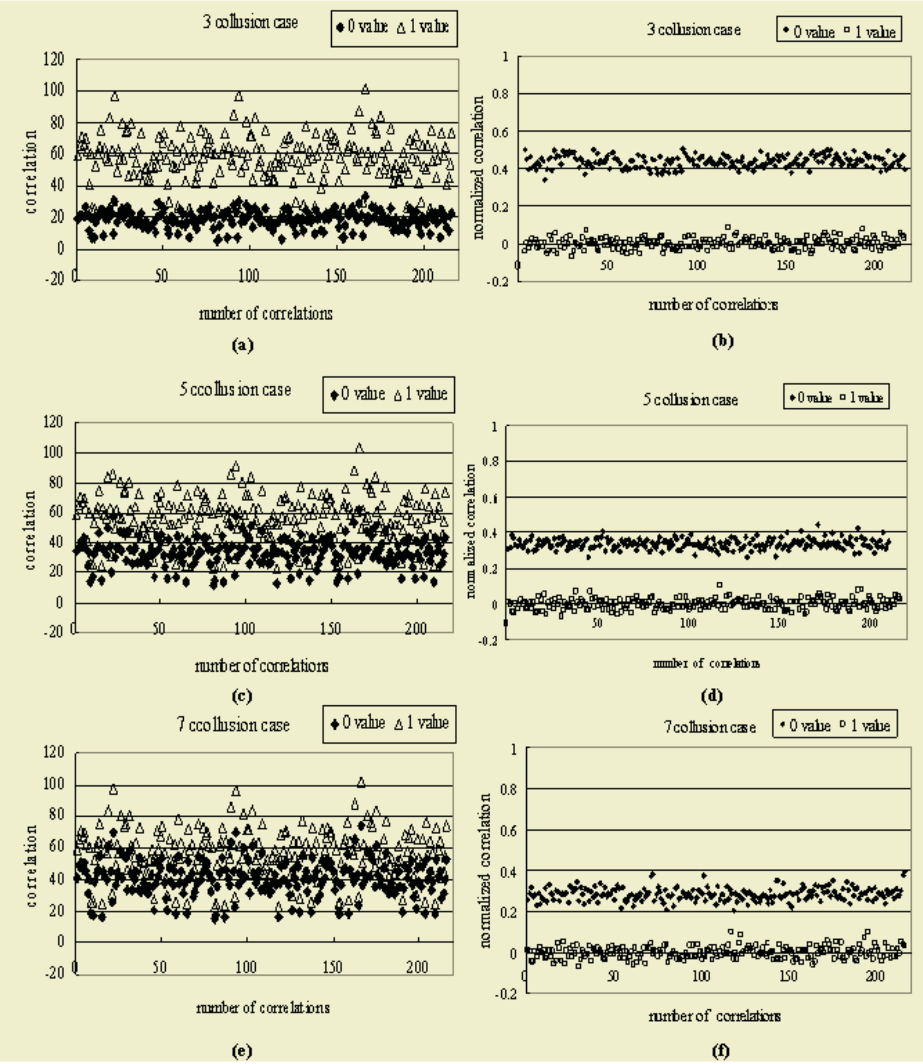


Fig. 3. Correlation value separation results both systems: detected correlation values for 0 bit (black colored shape) and 1 bit (white colored shape). (a), (c) and (e) are result of legacy systems and (b), (d) and (f) are of the proposed system

5 Conclusions

The anti-collusion fingerprint code scheme has a property of the colluder tracing function based on a marking assumption concept. The marking assumption, however, does not work well in multimedia data because the detection scheme of fingerprint signals works with correlation-based method and the results are not always agree with its theoretical result. This disagreement cause wrong detections of innocent users, which is the most serious problem in fingerprinting applications. To get rid of this trouble, we proposed a zero-based code modulation method which fully reflects running mechanisms of marking assumption and anti-collusion codes. We modulated and embedded only 0 bit information of binary codes so that correlation results for bit 0 and 1 could be separated as maximally as possible. Through code bit separation experiments, we convinced our proposed method is effective to classify two groups of correlation points even though the worst cases of averaging attacks. From video data experiments, we confirmed that the bit separation property was maintained and the detector could trace colluders successfully in the corrupted contents.

References

1. Junseok Lee, Seong O.H.: A DRM Framework for Distributing Digital Contents through the Internet, ETRI Journal, 2003, pp. 423-436.
2. D. Boneh, J. Shaw: Collusion-secure fingerprinting for digital data, IEEE Trans. On Information Theory, Vol. 44. (1998) 1897-1905
3. W. Trappe, M. Wu, Zhen Wang, K.J.R. Liu: Anti-collusion Fingerprinting for Multimedia, IEEE Trans. On Signal Processing, Vol. 51. (2003) 1069-1087
4. Z. Jane Wang, Min Wu, Hong Zhao, K. J. Liu and Wade Trappe: Resistance of orthogonal Gaussian fingerprinting to collusion attacks, Proc. of ICASSP, pp 724-727, Apr. 2003
5. C. J. Colbourn, J. H. Dinitz: The CRC Handbook of Combinatorial Designs, CRC (1996)
6. Willard H. Clatworthy: Tables of two-associate-class partially balanced designs, National Bureau of Standards Washington D.C. U.S. (1973)
7. I. J. Cox, M.L. Miller, J.A. Broom: Digital Watermarking, Morgan Kaufmann Publishers: San Francisco CA (2002)

Studies on Method for Measuring Human Feelings

Taki Kanda

Bunri University of Hospitality, Department of Service Management,
311-1 Kashiwabarashinden Shinogawara, Sayama,
Saitama 350-1336, Japan

Abstract. In this paper it is studied how to measure human feelings. Since human feelings are not simple and very complicated and it is not easy to evaluate them. Generally human feelings are considered nonlinear and methods for paired comparison are useful to deal with nonlinear problems. For the reason it is studied how to use paired comparison for measuring human feelings. Among many methods for paired comparison Case V of Thurston's method is perceived. For the measurement obtained by Case V of Thurston's method for paired comparison is useful because the units of the measurement are the standard deviations of the standard normal distributions Here it is discussed to measure human feelings using paired comparison.

1 Introduction

Methods of paired comparisons are useful for measuring human sensations or feelings. Measuring human sensations is a main concern of the study of Psychophysics created by Fechner in 1860. Thurstone's method of paired comparisons, which is a typical method of paired comparisons, was proposed as a method of psychophysical analysis and developed by Mosteller. Thurstone's method is a series of methods from questionnaire survey to statistical analysis for the analysis of human sensations and has been developed mainly for sensory tests as are Scheffe's method and Bradley's method. Besides these methods for paired comparisons, there are methods using paired comparisons to measure human subjective evaluation, which start with paired comparisons and end up solving eigenvalue problems. One is Guttman's method and another is AHP (Analytic Hierarchy Process), proposed by Saaty. Guttman's method was developed with the aim of determining the order of priority to demobilize American soldiers after the Second World War. AHP was studied as a method to properly combine subjective judgment with a systematic approach and is now widely adopted in Europe, U.S.A. and other areas in various fields, such as economic problems, management problems, energy problems, policy decision, and city planning. Among the above mentioned methods for paired comparisons, Thurstone's method has a long history and is a typical method in Psychometrics, useful for measuring not only sensations, but also feelings. Taking the feelings about food, for example, it is possible to measure our feelings about food or to classify menus on home dining tables according to human feelings using the method and related techniques.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1115–1121, 2005.

2 Difference of Sensations and Feelings

Human sensations and perceptions are considered separately in psychology, and human sensations and feelings should also be considered separately. We feel something about objects in the manner shown in Figure 1: first, we are conscious of the existence of objects through sensory organs, perceive them based on our knowledge, recognize them through our experience or learning, and feel something about them variously mixing our preference, emotion, or sentiment. Methods for paired comparisons have been used for measuring human sensations and when sensations are measured by using paired comparisons, pairs of objects are compared. On the other hand, for measuring human feelings, words are useful to make comparisons. In order to measure our feelings for food, the framework of human meal behavior should be defined.

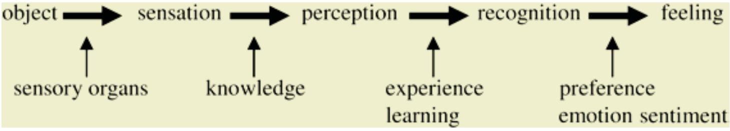


Fig. 1. Process of having feelings toward objects

3 Framework of Human Meal Behavior

We must take meals to be alive and active. Regardless of our situation or interest in taking meals we cannot survive without eating. This is the reason for behavior relating to eating conveniently at meals. □e also eat not only for the need to be alive but for enjoyment. Enjoying meals means to search for good taste or to put much importance on the culture of taking meals. This is the reason for the earnestness of meals and meal behavior. Moreover food has the health-related function such that eating certain foods is good for our health and eating others is bad for our health. In this manner, we are concerned about our bodies when eating. This is the reason for behavior relating to eating healthily at meals. In the towns in Japan, for example, many fast food shops or convenience stores can be found and it follows that there are quite a few people who want to finish their meals conveniently. On the other hand, when fast food shops first appeared in Italy in the 1980s, people felt that these shops would ruin Italian food culture and in 1986 at a village BRA in Italy, the slow food movement arose against fast food. In 1989, the slow food convivium (President: Carlo Petrini) was established with the aim of preserving Italian food culture and now has more than seven hundred thousand members from all over the world. Slow food convivium is now established at several places in Japan (Tokyo. Miyagi, Yamagata etc.), too. It can naturally be said that there are many people who consider meals important and serious in terms of their attitude to life, that is, they take meals earnestly. In addition to human "convenience-oriented meal behavior" and "earnestness-oriented meal behavior", human "health-oriented behavior" is also obvious because we have been seeing various foods known as "health food" or "natural food" sold and now many kinds of functional food can be found. For the above reasons the framework of human meal behavior is defined as "convenience", "health", "earnestness".

4 Thurstone's Method of Paired Comparisons

In Thurstone's method of paired comparisons there are 5 ways - case I, case II, case III, case IV and case V depending on assumptions about the distribution of human sensations. Since case V gives us scales, the units of which are the standard deviations of the standard normal distribution, the scales are very helpful for us to compare human sensations or feelings among many objects. In case V of Thurstone's method, however, it is assumed that human sensations are normally distributed and the standard deviations of each distribution are equal. It is therefore necessary to find out whether scales obtained by paired comparisons fit the assumption of case V of Thurstone's method.

5 Experiment by Paired Comparisons

The results are shown an experiment on human feelings about food by paired comparisons conducted with the stimuli being the words "convenience", "health" and "earnestness". In the experiment, the subjects, 135 University students, were equally divided into 3 groups, that is, each group consisted of 45 subjects and were asked three different questions, depending on group, with respect to 200 menus on home dining tables. Specifically, one group was asked whether each menu was convenience oriented or health oriented, another group was asked whether each menu was health oriented or earnestness oriented, and the third group was asked whether each menu was earnestness oriented or convenience oriented.

6 Test of Goodness of Fit

As for the results obtained, we need to test the goodness of fit to the model assumed in Case V of Thurstone's method of paired comparisons. To do this we obtain the estimated proportions EP_{ch}, EP_{he} and EP_{ec} corresponding to the differences of the convenience-oriented and health-oriented average values, the health-oriented and earnestness-oriented average values, and the earnestness-oriented and convenience-oriented average values respectively. We also obtain the true proportions TP_{ch}, TP_{he} and TP_{ec} for EP_{ch}, EP_{he} and EP_{ec} respectively. Now for each proportion an arcsine transformation is made as follows:

$$\begin{aligned} E\theta_{ch} &= \arcsin \sqrt{EP_{ch}} \\ E\theta_{he} &= \arcsin \sqrt{EP_{he}} \\ E\theta_{ec} &= \arcsin \sqrt{EP_{ec}} \\ T\theta_{ch} &= \arcsin \sqrt{TP_{ch}} \\ T\theta_{he} &= \arcsin \sqrt{TP_{he}} \\ T\theta_{ec} &= \arcsin \sqrt{TP_{ec}} \end{aligned} \qquad (1)$$

We calculate the test statistic

$$\chi_0{}^2 = \frac{n[(E\theta_{ch} - T\theta_{ch})^2 + (E\theta_{he} - T\theta_{he})^2 + (E\theta_{ec} - T\theta_{ec})^2]}{821} \qquad (2)$$

and obtain the critical value $\chi^2(df_1, \alpha)$ at the significance level α for degrees of freedom

$$df_1 = \frac{1}{2}(n-1)(n-2) \tag{3}$$

where n is the number of stimuli. If $\chi_0{}^2 < \chi^2(df_1, \alpha)$ is satisfied, it is determined that feelings fit the assumption of Thurstone's method .

The fitness of feelings to Thurstone's assumption has been tested for 200 menus on home dining tables including fried fish dressed with liquid starch and seafood and vegetables dressed with liquid starch. In this case stimuli are menus with convenience-oriented feelings, health-oriented feelings and earnestness-oriented feelings. Hence there are three stimuli, n=3. Substituting this in (3) we obtain $df_1 = 1$. Thus the critical value at 5% level of significance comes out

$$\chi^2(1, 0.05) = 3.84 \; . \tag{4}$$

The test statistic $\chi_0{}^2$ was obtained from (2) to be 1.93, which is less than the critical value 3.84 for fried fish dressed with liquid starch and 4.84, which is not less than the critical value 3.84 for seafood and vegetables dressed with liquid starch. It is therefore statistically determined that the observed feelings fit the assumption of Case V of Thurstone's method for fried fish dressed with liquid starch but the observed feelings do not fit the assumption for seafood and vegetables dressed with liquid starch. The results were obtained for 200 menus on home dining tables as follows: the observed feelings fit the assumption for 51 menus and do not fit the assumption for 149 menus. These results show that it is considered impossible for evaluation of feelings to use the values obtained from the experiment. It is therefore necessary to consider another method to analyze human meal feelings. Thus a method is considered to classify menus on home dining tables according to human feelings using the method of paired comparisons.

7 Agreement of Feelings

In order to classify menus on home dining tables using the data obtained from the experiment on human feelings, it is necessary to confirm the agreement of the observed feelings among many subjects. To do this, coefficient of agreement is used.

Let k be the number of subjects, n be the number of stimuli and x_{ij} be the number of subjects who choose stimulus i in the comparison of stimuli i and j. Coefficient of agreement is given by

$$u = \frac{2[{}_kC_2 \times {}_nC_2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} x_{ij}(k - x_{ij})]}{{}_kC_2 \times {}_nC_2} - 1 \tag{5}$$

where $-1 \le u \le 1$. The values obtained by (5) show how the observed feelings agree among subjects. To test the significance of agreement on feelings, the chi-square test is conducted. To do this, first, we calculate the test statistic

$$\chi_0{}^2 = \frac{4}{k-2}[{}_kC_2 \times {}_nC_2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} x_{ij}(k - x_{ij}) - \frac{kn(k-1)(k-3)(n-1)}{8(k-2)}] \tag{6}$$

and next we obtain the degrees of freedom. The value of the degrees of freedom is given by

$$df_2 = \frac{kn(k-1)(n-1)}{2(k-2)^2} \cdot \tag{7}$$

If $\chi_0^2 \geq \chi^2(df_2, \alpha)$ is satisfied, it is determined that feelings agree at the significance level α . Here $k = 45$, $n = 3$ hence we have $df_2 = 3.21$. The critical value $\chi^2(df_2, 0.05)$ at 5% level of significance comes out as

$$\begin{aligned} &7.81 = \chi^2(3, 0.05) < \chi^2(df_2, 0.05) < \chi^2(4, 0.05) = 9.50 \\ &\therefore\ 7.81 < \chi^2(df_2, 0.05) < 9.50 \ \cdot \end{aligned} \tag{8}$$

The test statistic χ_0^2 is calculated by (6). It is 4.98 for fried fish dressed with liquid starch and 5.91 for seafood and vegetables dressed with liquid starch. Both values of χ_0^2 for fried fish dressed with liquid starch and seafood and vegetables dressed with liquid starch are less than 7.81 hence they are less than the critical value $\chi^2(df_2, 0.05)$. It is therefore determined that the observed feelings for fried fish dressed with liquid starch and seafood and vegetables dressed with liquid starch do not agree among subjects. The agreement of feelings was tested with respect to 200 menus on home dining tables and it has been concluded that the observed feelings for 35 menus (17.5%) agree and those for 165 menus (82.5%) do not agree among subjects. Tables 5 and 6 show 15 menus in which feelings do not agree (Table 1) and feelings agree (Table 2) as examples. In these tables, the values of the agreement of coefficient *u* and test statistic χ_0^2 for each menu are also shown.

Table 1. The menus in which feelings do not agree

Menus	u	χ_0^2
Slices of seafood	-0.02	0.33
Pasta salad · Salad noodles ·Chinese noodles salad	-0.01	2.00
Eggs fried in a pan	0.01	4.05
Cooked rice and red beans	0.01	4.79
Cheese	0.01	4.98
Tea	0.01	5.17
Broiled vegetables	0.02	5.54
Chinese soap	0.02	6.10
Bread	0.03	6.84
Broiled shellfish	0.03	7.03
Pickled ume	0.03	7.03
Vegetables boiled down in soy	0.03	7.40
Boiled fish paste	0.03	7.40
Minced fish	0.03	7.77
Boiled plain ·chowder	0.03	7.77

Table 2. The menus in which feelings agree

Menus	u	χ_0^2
Toast	0.04	9.70
Stick salad	0.07	12.61
Meuniere of fish	0.09	15.96
Rice ball	0.12	19.49
Mixed rice	0.13	20.79
Rice fried in a hot pan	0.14	22.10
Fish teriyaki	0.18	27.49
Kinds of bean curd	0.20	31.03
Mixed salad	0.24	35.86
Steak	0.27	40.51
Green fruit	0.27	40.51
Gratin	0.29	43.86
Bacon	0.30	44.42
Laver	0.30	44.79
Dried vegetables and seafood	0.38	56.33

8 Classification of Menus

Menus on home dining tables are classified into four groups: a group of convenience-oriented menus, a group of health-oriented menus, a group of earnestness-oriented menus, and a group of non-oriented menus. The menus for which it is statistically determined that the observed feelings do not agree among subjects are put into a group of non-oriented menus. Therefore fried fish dressed with liquid starch and seafood and vegetables dressed with liquid starch, taken as examples of menus on home dining tables from 200 menus, belong to a group of non-oriented menus. There are 35 menus among 200 menus in which feelings do not agree and those menus belong to a group of non-oriented menus. The menus listed in Table 1 therefore are examples of non-oriented menus. As for the menus in which feelings agree, if the scale of the convenience-oriented feeling is the largest compared with those of two other feelings, the menu is put into the group of convenience-oriented menus, if the scale of the health-oriented feeling is the largest, the menu is put into the group of health-oriented menus, and if the scale of the earnestness-oriented feeling is the largest, the menu is put into the group of earnestness-oriented menus. Table 3 shows examples of the menus in which it is statistically determined that feelings agree and their scales for convenience-oriented feeling, health-oriented feeling and earnestness-oriented feeling. In terms of Table 3, this means that corned beef turns out to be a convenience-oriented menu because the scale for the convenience-oriented feeling is the largest, yogurt turns out to be a health-oriented menu because the scale for the health-oriented feeling is the largest, and spagetti turns out to be an earnestness-oriented menu because the scale for the earnestness-oriented feeling is the largest.

Table 3. Scales of menus

Menu	Convenience	Health	Earnestness
Corned beef	1.499	-1.442	-0.057
Yogurt	-0.615	1.923	-1.308
Spagetti	0.304	-1.513	1.209

9 Final Remarks

The measurement of human sensations or feelings is helpful in various fields and methods of paired comparisons are useful to do this. Studies on methods of paired comparisons have a long history and many methods have been proposed. In order to use methods of paired comparisons effectively, it is important to determine what to measure, sensations or feelings etc., choose appropriate methods from the many methods of paired comparisons, and consider how to use them according to the problems at hand.

References

1. Thurstone, L.L.: Psychophysical Analysis. American Journal of Psychology38 (1927)368-389.
2. Thurstone, L.L.: A Low of Comparative Judgment. Psychological Review34 (1927)273-286.
3. Mosteller F.: Remarks on the Method of Paired Comparisons: III. Test of Significance for Paired Comparisons When Equal Standard Deviations and Equal Correlations are Assumed. Psychometrika16(2)(1951) 207-218

Data Mining Method from Text Database*

Masahiro Kawano[1], Junzo Watada[1], and Takayuki Kawaura[2]

[1] Waseda University I.P.S.
[2] IICLO

Abstract. Recently, various types of data are expected to get in information processing according to multi-media technology. Especially, linguistic data are employed in fuzzy systems as well as fuzzy numerical values.
In this paper we propose a text minig method based on fuzzy quantification model. In the process of text mining, we will pursue the following steps: 1) Sentences included in a text in Japanese are broken down into words. 2) It is possible to realize common understanding using fuzzy thesaurus that enables us to translate words into synonyms or into upper concepts.
In this paper, we employ the method to translate words using Chinese characters or continuous letters of Katakana more then one katakana letter (Japanese alphabet letter) into keywords. The method realizes the high speed of processing without any dictionary for separating words.
Fuzzy multivariate analysis is employed to analyze such processed data and to abstract a latent mutual related structure under the data. In other words, we abstract the knowledge from the given text data.
At the end we apply the method to mining the text information of libraries and Web pages distributed over a web network and discussing about the application to Kansei engineering.

Keywords: Text mining, fuzzy quantification analysis, library data.

1 Introduction

Recently, various types of data are expected to get in information processing according to multi-media technology. Especially, linguistic data are employed in fuzzy systems as well as fuzzy numerical values.

In this paper we propose a text minig method based on fuzzy quantification model. In the process of text mining, we will pursue the following steps: 1) Sentences included in a text in Japanese are broken down into words. 2) It is possible to realize common understanding using fuzzy thesaurus that enables us to translate words into synonyms or into upper concepts.

[*] This paper is submitted to KES2005 "Ninth International Conference on Knowledge-Based Intelligence Information & Engineering Systems" held on September 14-16, 2005, at Melbourne, Australia.
All inquires can be sent to Junzo Watada, Professor / Waseda University / Graduate School of Information, Production and Systems / 2-2, Hibikino Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan / e-mail: junzow@osb.att.ne.jp

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1122–1128, 2005.

In this paper, we employ the method to translate words using Chinese characters or continuous letters of Katakana more then one katakana letter (Japanese alphabet letter) into keywords. The method realizes the high speed of processing without any dictionary for separating words.

Fuzzy multivariate analysis is employed to analyze such processed data and to abstract a latent mutual related structure under the data. In other words, we abstract the knowledge from the given text data.

At the end we apply the method to mining the text information of libraries and Web pages distributed over a web network and discussing about the application to Kansei engineering.

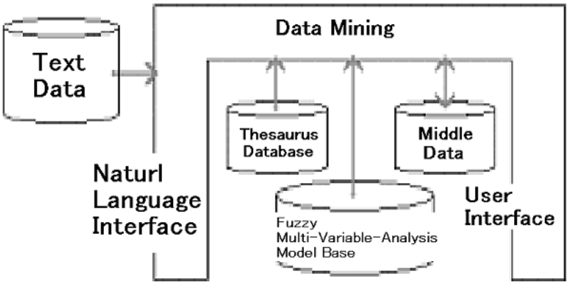

Text Data
Data Mining
Thesaurus Database
Middle Data
Naturl Language Interface
Fuzzy Multi-Variable-Analysis Model Base
User Interface

Fig. 1. Text Mining

2 Text Mining

In statistics or statistical data analysis, data are intentionally gathered to answer some questions, today data are hugely pilled out of the results of daily works. It is not easy to properly build appropriate hypothesis for such accumulated huge data. Therefore, data mining can enables us to discover some unsuspected patterns in a large-scale data-base. The method is widely employed in the real world. The characteristics of data mining can be listed up as follows:

1. the target is on a large-scale data-base,
2. generally the unintended accumulated data is analyzed,
3. the new types or new patterns are discovered, and
4. The collaboration or co-works between human and computers are the most important issue.[1]

In this paper, we will discuss about text mining. The most difference between conventional data mining and text mining is to deal with descriptive or linguistic data including sentences as well as with numerical data.

In text mining it is usual to pre-process text data such as separation of words or pigeonholing data using a thesaurus or into keywords in terms of natural language processing. The resulted set of words will be employed in text mining. Figure 1 shows the process of applications.

At the end, we will discuss text mining to text data of Kansei Engineering.

2.1 Fuzzy Thesaurus

It is essential to process text data into an appropriate form for analysis. When various words are used for the same meaning, the thesaurus enables us to unify several words into the same keyword. For example, in the case where “a mobile” and “a mobile phone” are used in a text, we should unify these words into one common keyword. And when several words can be included in the same category, we may use the upper concept which can unify words using an upper concept. For example, when a dog and a cat are used in the text, we can translate such words into a pet.

In order to enable this task we should build a thesaurus previously. In this paper, we build the thesaurus using fuzzy membership functions. The fuzzy thesaurus can assign one keyword to several related words and enables us to handle text data from various meaning and purposes.

2.1.1 Classification of a Text into Words

As mentioned above, it is required to separate a text into words in the previous task in mining from text data. It consits of the system which dynamically generates a set of words for a web [2].

In the system, it is not necessary to employ a dictionary data base previously developed. Chinese characters and Katakana (Japanese alphabets) are used for discriminations of words. Chinese characters and katakana are identified using JIS code and Katakana is separated and also more than one Chinese character is recognized as a keyword. Using this method the wide range of texts is analyzable. Even newly used words can be recognized as a word without any dictionary data base. That is, unknown words can be recognized as a keyword.

2.1.2 Building Fuzzy Thesaurus

In this paper, a fuzzy thesaurus based on fuzzy relations among words is employed in clustering obtained words into a thesaurus data base. There are many papers reported already. It is possible to handle words, letters and sentences as in real situation.

The strength of relation between words is expressed in the fuzzy thesaurus.

For example, when there are three persons who have interest in Chiwawa, They can be defined as having interest in a dog with 0.65, a pet with 0.9 and an animal with 0.55.

2.1.3 Usage of Fuzzy Thesaurus

Generally, a thesaurus is employed to translate keywords. For example, in the case that plural persons express their fondness of a dog, a cat or other pets, the word placed upper than these words is translated from these words and these people is categorized into the group of pet lovers.

In this paper, a fuzzy thesaurus is treated such as that new keywords will be automatically added as well as the translation will be done using the thesaurus.

The method strengthens mining techniques to abstract the latent structure or latent features which can not be posed out on the surface of the data.

In the method sometime the coincidence happens that an unacceptable result occurs. In order to remove such a possibility, we employ a membership function into the above-mentioned translation. The fuzzy relation enables us to analyze the data precisely.

3 Application of Text Mining

3.1 Mining from Library Data Bases

In this paper, we will explain the method how to mine Library Data Bases consisting of descriptions of reading books.

The data base includes age, sex and grade of school about a person who used a library as well as the classification code of a borrowed book, title, author, recommended grade of school, publisher, total page, year of the first edition, ISBN code, NDC, keywords and abstract.

3.2 Data Mining from Web Data

In this paper, we asked the sample to answer descriptively questions in the descriptive sentences instead of alternatives.

Generally, we employ questionnaires to gather data, opinions and so on. The questionnaires should be previously built so as to obtain some information which satisfies the objective of its survey. This requires us to narrow our question or answering style in many cases. Therefore, these questionnaires employ question style limited its answers into Yes/No alternatives in order to make its analysis easier. On the other hand, it is not appropriate to limit the answering style into alternative style if we expect more flexible or sufficient information on the objective. We intend to analyze such obtained data written in a texts descriptive instead of Yes/No questionnaires.

Conventionality, one of questionnaires are formed that ask a sample to select alternatives for a question. This type of questionnaires limits the selection of a sample when an appropriate question is asked. As a result, the in In this case we gathered the data using Web questionnaires. That is, we will analyze data gathered from the Web board which is provided to obtain some purposes in a library.

The Web board is employed so that writers describe their own opinions in descriptive style when the provider of the board requires their theme on the board.

Therefore, the Web board has less restrict or limitation in answering or giving own opinions. It is much appropriate to gather sufficient information from people.

The defects of the method are on that the Web board is open toward to undecided people and therefore, the answering is done toward such people. The reliability on the content given on board will be decreased.

It is possible to take a keyword for the opinion of the person and gather the page as an opinion of the answering person when the page includes some objective keyword.

Even in the case, it is required to understand the sentences in order that the text data can be effectively employed. For this purpose, it is essential to understand verbs in the text.

4 Application of Text Mining to Kansei Engineering

In this section, we propose the method to Kansei Engineering. That is, we will explain how the knowledge acquired by text mining from Kansei information can be employed in marketing or sales or production.

Kansei information is essential in modern product development. In other words, in developing a new product, the kansei information is very important relating to design and functions.

The objective of Kansei engineering is a technology which translate the image or kansei which consumer used to have in its usage to the design level that realizes the concrete subject [5]. Such psychological measurement method in evaluation experiment of Kanasei Engineering used to be pursued in terms of words or verbal expression,

Text mining is a method to acquire knowledge from descriptive sentence or words. Therefore, the text mining is appropriate to Kansei Engineering.

The procedure of Knasei Engineering is executed as follows:

- (1) abstract Kansei words relating to surveyed subject,
- (2) evaluate the words based on Semantic differential method or Quantification methods,

where Kansei words are generally adjective words which express some feeling such as “pretty”, “cozy” and so on.

Conventionally Kansei words are gathered from consumers of a product or people of retailing that evaluate words appropriate to the product or from a catalog or a dictionary. Such time consuming tasks are required.

4.1 Decision Making of Selecting Appropriate Books for a Library from the Results of Library Data

There are differences of interesting genre among school grades of children. The result shows what kind of journal children are interested in among different grades.

Children in a kindergarten has strong interest in books on a kindergarten or animals. Male pupils of an elementary school or of a junior high school are interested in vehicles and boys in higher grade like to read are books on a computer. It is junior high school pupils who interested in ghost stories and fortune telling.

Such results can be employed in selecting new books for a library and accelerates the usage or development of libraries.

4.2 Gathering Kansei Information from Web for Development of Products

Let us discuss about the results of above-mentioned text mining from Web. Since it is essential to understand the meaning of sentences in the case of such data, we did not separate a sentence into words. Nevertheless, if we can find such symbols as expressing the end of a sentence, we separate these sentences into a single one.

When the objective of analysis is clear as in such a board in web, we can first look for words used for preference or feeling expression. After then, we check the duplication of negation in a sentence such as "it is not, that, not." This is done to understand its meaning clearly and properly. For example, even though we find a keyword "dislike" or "hate", the real meaning does not express it but "like" or "love" when the negation is follows in the sentence. The meaning is rather weaker than the strait expression although it means positive. Therefore, we evaluate the grade of the meaning as well as distinguishes the positive and negative. For example, the positive is grade 1 and the negative grade 0, but the expression with both dislike and negation is grade 0.5.

In this case, we can express the grade or portion of the preference to the objective under the survey. And also, we can distinguish words before or after the affirmative or negative keywords in order to understand what kind of products or portion they have preferences.

5 Concluding Remarks

The objective of the paper is to provide a method for text mining on the base of a fuzzy thesaurus method and fuzzy relation and to provide several real applications. The results of the applications show the effectiveness of our method.

In order to realize the method, we should build the huge dictionary of a fuzzy thesaurus. In this paper, we constructed the thesaurus from the past experience. It required the huge computation to computing time and costs. Therefore, it is issue how to build the real huge thesaurus.

Acknowledgement

We would like to express our acknowledge to Mr. Nobuharu Mochida, HIMAC.JP who assist building a system.

References

1. Keisuke Aoki, Junzo Watada, "Data mining method from text database bsed on fuzzy quantification analysis," Proceedings, IEEE SMC 2004, on 2004.10.1, Hague, Netherlands, pp.6472-6478 (2004).
2. S. Mochida, "Dynamic knowledge set generation system that used Web technology," Proceedings of Kyushu and Yamaguchi Branch, Bio-Medical Fuzzy Systems Association, 2003.8.23, in Japanese.

3. M. Nagamachi, *Basic Study of Kansei Engineering and Application*, Kaibundou publication Ltd., 1993, in Japanese.
4. Y. Nakamori, *Fuzzy Quantification Analysis for Kansei Data Analysis and Sensibility Information Processing*, Morikita Publishing, 2000, in Japanese.
5. Watada, J. "Fuzzy Quantification Theory", Chapter 6, Edited by T. Terano, K. Asai and M. Sugeno, *Fuzzy Systems Theory and Its Applications*, pp. 101-123, Academic Press, 1992.
6. Zadeh, L. A., "Fuzzy Sets,"*Information and Control*, 8, pp.338-353 (1965).
7. Zadeh,L. A., Probability Measures of Fuzzy Events, *Journal of Mathematical Analysis and Applications*, 23, pp.421-427 (1968).

The Air Pollution Constraints Considered Best Generation Mix Using Fuzzy Linear Programming

Jaeseok Choi[1], TrungTinh Tran[1], Jungji Kwon[1],
Sangsik Lee[1], and Abdurrahim El-keib[2]

[1] Department of Electrical Engineering
Gyeongsang National University
900, Gazwa-dong, Chinju GN, Korea 660-701
jschoi@nongae.gsnu.ac.kr
tttinh73@hotmail.com
[2] Department of Electrical and Computer Engineering
University of Alabama
Tuscaloosa, AL, USA
El-keib@coe.eng.ua.edu

Abstract. A new approach considering SO_x, NO_X and CO_2 air pollution constraints in the long-term generation mix with multi-criteria is proposed under uncertain circumstances. Specially, CO2 emission of electricity system industry has over thirty percent of total emission quantity in the world. The CO2 emission in coal power plant competitive with nuclear power plant is very severe. The air pollution in coal is requiring LNG or new environmental type generation system (wind power, solar power tidal power et al.) instead of coal power plant, despite the new generation systems ask for very high construction cost. A characteristic feature of the presented approach is what effects is the air pollution constraints in long term best generation mix. The fuzzy linear programming is used for analyzing ambiguity in this study. A characteristic feature of the presented approach is that not only fuzziness in fuel and construction cost, load growth, reliability and air pollution but also many constraints of generation mix can easily be taken into account by using fuzzy linear programming. The proposed method accommodates the operation of pumped-storage generator. The effectiveness of the proposed approach is demonstrated by applying to the best generation mix problem of KEPCO-system, which contains nuclear, coal, LNG, oil and pumped-storage hydro plant in multi-years.

1 Introduction

There is a global trend towards liberalization and privatization of the electricity supply industry. This is coupled with growing environmental awareness and increasing prospects ratification of the Kyoto Protocol.[1] Electricity is the indispensable form of energy in modern societies. Its demand has been increasing more and more quantity, quality and reliable at minimize production cost. The restructuring of electricity market has been moving from monopolistic to competitive that split generation, transmission and distribution sector in power system into GENCO, TRANSCO and DISCO respectively.[2] The goal of the GENCO is to product electricity to meet growing load demand for a reasonable price, environmental cleanness, and a reliable

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1129–1142, 2005.

quality electric energy source. A huge budget associated with the construction, difficulty for a place of plant, unstable security of fuels, uncertainty within long-term load forecasting and the constraint of the air pollution, the generation planning is getting very complicated.[3-5] Under these unstable circumstances, decision makers are falling into a difficult situation in these days. It is quite important that not only the economic aspect and the reliability but also air pollution in generation system and the uncertainty of the load are considered for the generation mix.[6] And so, planners are supposed to propose decision makers some methods to analyze the energy and power system on the basis of the relevant data for their dynamic and reasonable decision making. Several simple techniques are available for solving this generation mix under the uncertainties.[7-9] Electricity utilization is environmentally benign and as a form of energy carrier electricity is clean and safe. It causes no pollution or environmental emissions at the point of end user. But, electricity production can cause local and regional environmental impact and may also have long-lasting detrimental global consequences. Some of these impacts like the emissions of sulphur dioxide (SO_2), nitrogen oxides (NO_X) and solid particulates, which all have detrimental air quality implications, can be controlled by investing in technology and abatement facilities. These measures can be control and reduce significantly such emissions. Carbon dioxide (CO_2), which is the main gas suspected of causing global warming (greenhouse effect), is far more difficult and expensive to control.[1]

In this paper, a new approach using new fuzzy linear programming is proposed for the long-term generation mix with multi-criteria considering air pollution constraints, which are SO_x, NO_X and, furthermore, CO_2 emission limitation, under the uncertain circumstances. Essentially the theory consists of the imposition of the framework of a fuzzy decision on the linear programming concept. The method can accommodate a linear shape of the membership function as well as the operation of the pumped-storage generator. Economics, reliabilities, air pollution and the uncertainties of the load are evaluated from the membership functions given at all states. Bellman-Zadeh's maximization decision process is used for searching optimal solution.[10-12]

2 Fuzzy Set Theory and Fuzzy Linear Programming

The fuzzy decision D resulting from q fuzzy goals $G_1, \ldots, G_q$ and p fuzzy constraints $C_1, \ldots, C_p$ is the intersection of them.[10,11]

$$D = \left(\bigcap_{i=1}^{q} G_i \right) \cap \left(\bigcap_{j=1}^{p} C_j \right) \tag{1}$$

And also its membership function μ_D resulting from fuzzy goals and constraints is defined by

$$\mu_D(x) = min\left[\underset{i=1\sim q}{min}\, \mu_{Gi}, \underset{j=1\sim p}{min}\, \mu_{Cj} \right] \tag{2}$$

The fuzzy mathematical programming problem consists of finding the maximum of the fuzzy decision D

$$\mu_D\left(x^*\right)= max\,\mu_D\left(x\right) \tag{3}$$

where *x** is the optimal decision solution.

The vector eq.(3) can be rewritten as the eq.(4)

$$\mu_D\left(x_1^*, x_2^*, \cdots, x_N^*\right)= \max_{x_1 \cdots x_N} \mu_D\left(x_1, x_2, \cdots, x_N\right) \tag{4}$$

If the decision maker wants to have a "crisp" decision proposal, it seems appropriate to suggest to him the dividend satisfaction level which has the highest degree of membership in the fuzzy set "decision". This decision in fuzzy environment has formally been defined by Bellman and Zadeh in their paper 1970.[11] The problem here is to maximize the satisfaction level, λ of the decision maker subject to fuzzy construction cost/budget. Therefore, eq.(4) can be formulated as eq.(5). Membership function that reflects uncertainties and ambiguities associated with construction cost is used. Fig. 1 shows the concept of fuzzy optimal decision making.

$$\begin{aligned} \lambda^* &= Max\ \mu_{D(x)} \\ &= Max\ \left\{Min\left[\mu_{G(x)}, \mu_{C(x)}\right]\right\} \\ &= \mu_{D\left(x^*\right)} \end{aligned} \tag{5}$$

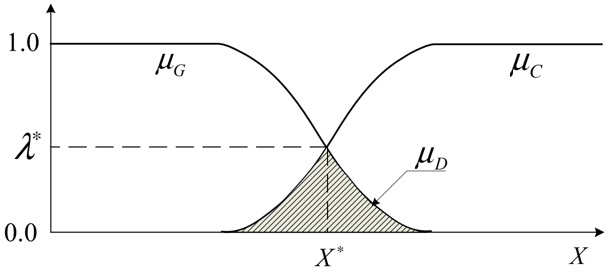


Fig. 1. Concept for optimal decision on fuzzy sets

We can rewrite the eq.(5) equivalently as eq.(6). Where, $\mu_{H(x)}$ is membership function of fuzzy set H(x) which presents fuzzy goals and fuzzy constraints sets together.

$$\begin{aligned} &\text{Maximize} \quad \lambda \\ &\text{Sub.} \quad \lambda \leq \mu_{H(x)_i}(x) \\ &\qquad\quad x \geq 0 \\ &\qquad\quad 0 \leq \lambda \leq 1 \end{aligned} \tag{6}$$

Under assumption that the $\mu_{H(x)}$ is linear function, we can formulate optimal problem for the fuzzy LP as eq.(7).

$$\begin{aligned} \text{Maximize} \quad & \lambda \\ \text{Sub.} \quad & \lambda\delta_i < H(x)_i < C_i + \delta_i \\ & x \geq 0 \\ & 0 \leq \lambda \leq 1 \end{aligned} \tag{7}$$

Where,

$$\mu H(x)_i = \begin{bmatrix} 1 & H(x)_i \\ \varepsilon[0,1] & C_i < H(x)_i < C_i + \delta_i \\ 0 & H(x)_i > C_i + \delta_i \end{bmatrix}$$

where, $\varepsilon[0,1] = 1 - \dfrac{H(x)_i - C_i}{\delta_i}$

3 The LP Formulation of Best Generation Mix

3.1 Problem Statement

This problem can be defined as to determine the generation mix under the following assumptions:

(1) The annual loads are given.
(2) The number of generator is not that of units but that of types.
(3) Nuclear power plants are able to perform load following.

The system for the proposed method can be modeled as shown in Fig. 2.

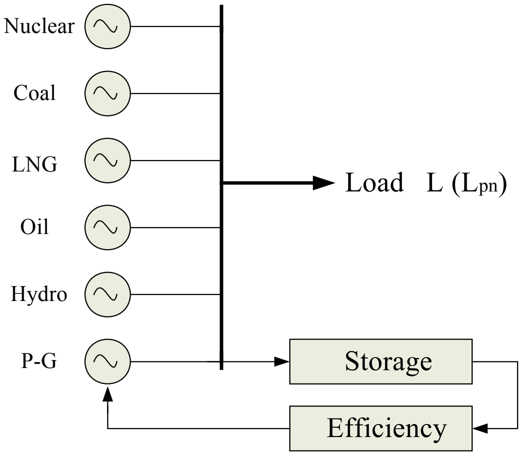


Fig. 2. A system model for the proposed method

In this study, it is assumed that the hydro generator construction is separately planned from that of the other kinds of generation units. In actual system, the basic resources, which are reserves, of the hydro power plants have limitation in the country. Therefore, the choice of hydro plant construction is not much and non-flexible. Under the assumption, the best generation mix problem is formulated as followings.

Objective Functions

The economic criterion:

The economic criterion in the best generation mix is minimization of the sum of the construction cost and the fuel cost as

$$\textit{Minimize } Z = \sum_{n=1}^{N}\sum_{i=1}^{NG} K_{cin}\, d_{in}\alpha_i\, \Delta x_{in} + \sum_{n=1}^{N}\sum_{i=1}^{NG} K_{fin} f_{in} y_{in} \tag{8}$$

$$= F(\Delta x_{in}, y_{in})$$

where, i: unit type number (1 for nuclear, 2 for coal, 3 for LNG, 4 for oil and 5 for pumped-storage generators are specified in this paper)

N: number of total study stage year

NG: number of unit type

K_{cin}= $((1+e_{ci})/(1+r))^{n\,\Delta T}$

K_{fin}= $((1+e_{fi})/(1+r))^{n\,\Delta T}$

e_{ci}: apparent escalation rate of construction materials of i-unit

e_{fi}: apparent escalation rate of fuel of i-unit

r: discount rate

ΔT: step size years of study years

d_{in}: construction cost of the i-unit in n year

f_{in}: marginal fuel cost of the i-unit in n year [won/MWh]

α_i: annual expenses rate of the i-unit

Δx_{in}: construction capacity of the i-unit in n year [MW]

y_{in}: generation capacity of the i-unit in n year [MWh]

The future investment budget cannot be guaranteed. Although the solution investment budget for the crisp problem is optimal, it is not necessarily practical. The secured budget is usually over or below the optimal solution cost of (1). This ambiguity gives the decision maker for expansion planning some flexibility (freedom). The objective function considering the ambiguity can be expressed based on fuzzy set theory and is called fuzzy goal function. The fuzzy goal function considering the decision maker's aspiration level for the total cost, Z_0, can be expressed by (9).

$$Z \lesssim Z_0 \tag{9}$$

Constraints

1) *Installed capacity constraint*

$$\sum_{i=1}^{NG} (x_{in} + \Delta x_{in}) \geq L^P{}_n (1+R_n) - HYD_n \qquad n=1 \sim N \tag{10}$$

where, R_n: supply reserve rate in n year. [p.u]

HYD_n: capacity of hydro generator in n year. It is assumed that the HYD_n is given in this study.

2) Energy constraint of demand

$$\sum_{i=1}^{NG} y_{in} \geq (\mathrm{L}^{\mathrm{P}}{}_{n} + \mathrm{L}^{\mathrm{B}}{}_{n}) \text{ x } 8760/2 + V_n - \mathrm{HYD}_n \text{ x } 8760 \text{ x } \mathrm{CF_H} \qquad n=1 \sim N \qquad (11)$$

where, $\mathrm{L}^{\mathrm{P}}{}_{n}$: peak load at *n* year
$\mathrm{L}^{\mathrm{B}}{}_{n}$: base load at *n* year
V_n: the added demand energy is caused by pumped-storage generator
$\mathrm{CF_H}$: average capacity factor of hydro generator

3) Production energy constraint of generation system

$$y_{in} \leq (x_{in} + \Delta x_{in}) \text{ x } 8760 \text{ x } \mathrm{CF}_i \qquad i=1 \sim NG, \quad n=1 \sim N \qquad (12)$$

where, CF_i: average capacity factor of the *i*-unit

4) Capacity constraint in initial year

$$x_{i1} = \mathrm{EX}_i \qquad i=1 \sim NG \qquad (13)$$

where, EX_i: capacity of the *i*- existing unit

5) Constraint of mutual relationship between existing generator capacity and new generator capacity (state equation)

$$x_{in+1} = x_{in} + \Delta x_{in+1} \qquad i=1 \sim NG, \quad n=1 \sim N \qquad (14)$$

6) Energy constraint of LNG thermal plant

$$y_{3n} \geq \mathrm{LEPmin}_n \qquad n=1 \sim N \qquad (15)$$

where, LEPmin_n *:* LNG thermal generator production energy for LNG minimum due to consumption in *n* year

7) Constraints of reservoir capacity of pumped-storage generator

$$V_n \leq (x_{5n} + \Delta x_{5n}) \text{ x } 8760 \text{ x } \mathrm{CF}_5 \qquad n=1 \sim N \qquad (16)$$

where, PSM: pumped-storage maximum possible time per day of pumped-storage generator

8) Energy balance constraints between pumped-storage and pumped-generator

$$y_{5n} = \eta_{\mathrm{pg}} \text{ x } V_n \qquad (17)$$

where, η_{pg} : efficiency of pumped-storage generator

9) Upper-lower constraints of total capacity of generators

$$X_i^{min} \leq x_{in} + \Delta x_{in} \leq x_i^{max} \qquad i = 1 \sim NG \qquad (18)$$

10) Upper-lower constraints of new unit capacity

$$\Delta X_{in}^{min} \leq \Delta x_{in} \leq \Delta X_{in}^{max} \qquad i = 1 \sim NG, \quad n = 1 \sim N \qquad (19)$$

11) CO_2 air pollution constraint

$$\sum_{i=1}^{NG} CO2_{in} \rho_i \, y_{in} \le \mathrm{CO2}_{\mathrm{MAXn}} \tag{20}$$

where, $CO2_{in}$: CO_2 density of the i- unit in n year [ppm/Ton]
$\mathrm{CO2}_{\mathrm{MAX}n}$: maximum quantity of CO_2 permitted in n year [Ton/yr]
ρ_i : fuel consumption rate of the i- unit [Ton/MWh]

12) SO_X air pollution constraint

$$\sum_{i=1}^{NG} SOX_{in} \rho_i \, y_{in} \le \mathrm{SOX}_{\mathrm{MAXn}} \tag{21}$$

where, SOX_{in} : SO_X density of the i- unit in n year [ppm/Ton]
$\mathrm{SOX}_{\mathrm{MAX}n}$: maximum quantity of SO_X permitted in n year [Ton/yr]

13) NO_X air pollution constrain

$$\sum_{i=1}^{NG} NOX_{in} \rho_i \, y_{in} \le \mathrm{NOX}_{\mathrm{MAXn}} \tag{22}$$

where, $NOX_{i,n}$: NO_X density of the i- unit in n year [ppm/Ton]
$\mathrm{NOX}_{\mathrm{MAX}n}$: maximum quantity of NO_X permitted in n year [Ton/yr]

Using eq.(7), the best generation mix problem considering uncertainty of budget can be formulated as followings. Where, δ_0, $\delta^{min,max}{}_{I}$ and $\delta'^{min,max}{}_{i}$ are the intervals of membership functions of decision maker's aspiration level for total cost, minimum, maximum capacities of new construction generators and total installed generators for study years respectively.

Maximize λ

Subject to $F(\Delta x_{in}, y_{in}) + \delta_0 \lambda \le Z_0 + \delta_0$

$$\sum_{i=1}^{NG} (x_{in} + \Delta x_{in}) \ge L^P{}_n (1+R_n) - \mathrm{HYD}_n$$

$$\sum_{i=1}^{NG} y_{in} \ge (L^P{}_n + L^B{}_n) \times 8760/2 + V_n - HYD_n \times 8760 \times \mathrm{CF_H}$$

$$y_{in} \le (x_{in} + \Delta x_{in}) \times 8760 \times \mathrm{CF}_i$$
$$y_{in} \le (x_{in} + \Delta x_{in}) \times 8760 \times \mathrm{CF}_i$$

$$x_{i1} = EX_i$$

$$x_{in+1} = x_{in} + \Delta x_{in+1}$$
$$y_{3n} \ge \mathrm{LEP}_{\mathrm{min}n}$$

$$V_n \le (x_{5n} + \Delta x_{5n}) \times 8760 \times \mathrm{CF}_5$$

$$y_{5n} = \eta_{pg} \text{ x } V_n \tag{23}$$

$$X_i^{\min} - \delta_i^{\min} \le x_{in} + \Delta x_{in} - \delta_i^{\min}\lambda$$

$$X_i^{\min} + x_{in} + \Delta\delta_i^{\min}\lambda \le X_i^{\max} + \delta_i^{\min}$$

$$\Delta X_i^{\min} - \delta_i^{'\min} \le \Delta\, x_{in} - \delta_i^{'\min}\lambda$$

$$\Delta\, x_{in} - \delta_i^{'\max}\lambda \le X_{in}^{\max} + \delta_i^{'\max}$$

$$\sum_{i=1}^{NG} CO2_{in}\rho_i y_{in} \le \text{CO2}_{\text{MAXn}}$$

$$\sum_{i=1}^{NG} SOX_{in}\rho_i y_{in} \le \text{SOX}_{\text{MAXn}}$$

$$\sum_{i=1}^{NG} NOX_{in}\rho_i y_{in} \le \text{NOX}_{\text{MAXn}}$$

4 Case Studies

In order to demonstrate the effectiveness of the proposed method the generation mix is performed on KEPCO system, which consists of 6 types generators over 20 years (2006-2026). It is assumed that the initial year is 2006. The step size of planning year is assumed as five years (ΔT=5). The maximum, minimum load and hydro capacity in standard years are listed in table 1. The characteristics and economic data are summarized in table 2 and table 3, respectively.

Table 1. Maximum load, minimum load, and hydro plant at standard years

Years	Peak load [MW]	Hydro [MW]	LEP (10^3Ton)
2006	48108	4300	--
2011	67340	5800	6500
2016	69111	6340	7000
2021	78111	7500	7500
2026	87880	8100	8000

Table 2. Maximum load, minimum load, and hydro plant at standard years

Gen. Type	Initial capacity [MW]	Fixed charge [10^5won/ kW]	A_{ER} of fixed charge [%]	Marginal fuel cost [Won/ kW]	A_{ER} of fuel cost [%]	Annual cost rate [%]	Ca-pacity factor [%]	Fuel cons rate [Ton/ MWh]	Density [ppm/Ton] CO_2,	NOx,	SO_X
Nucl.	16715	135.0	2	3.8	0	19	80	--			
Coal	17465	100.0	2	14.8	1	17	60	0.4030	700	450	500
LNG	14313	85.0	2	27.2	3	17	70	0.0500	450	200	300
Oil	4308	65.0	2	80.0	4	17	70	0.0234	600	200	100
P-G	1000	45.0	2	00.0	0	13	30	--			

(where: A_{ER} means apparent escalation rate and discount rate is assumed as 10%)

Table 3. Maximum and minimum of capacity per year and total capacity of construction of new generators ([MW])

Gen. Type	ΔX^{max}	ΔX^{min}	X^{max}	X^{min}
Nuclear	20,000	0	68,000	0
Coal	20,000	0	68,000	0
LNG	20,000	0	50,000	0
Oil	15,000	0	10,000	0
P-G	2,000	0	5000	0

Table 4. Maximum limitation of air pollutions (10^3 [Ton/yr])

Air pollution Type	2011	2016	2021	2026
CO_2	50	50	50	50
SO_X	50	50	50	50
NO_X	50	50	50	50

Results and Discussion

The simulated results in the three cases; a case not considering air pollution criterion and two cases considering air pollution criterion, are shown in Table 5. The proposed fuzzy set theory based on best generation mixes, which are simulated under considering air pollution criterion are different from the generation mix using conventional method. The results yield that the mix of nuclear power plants is increasing and that of coal power plants is decreasing.

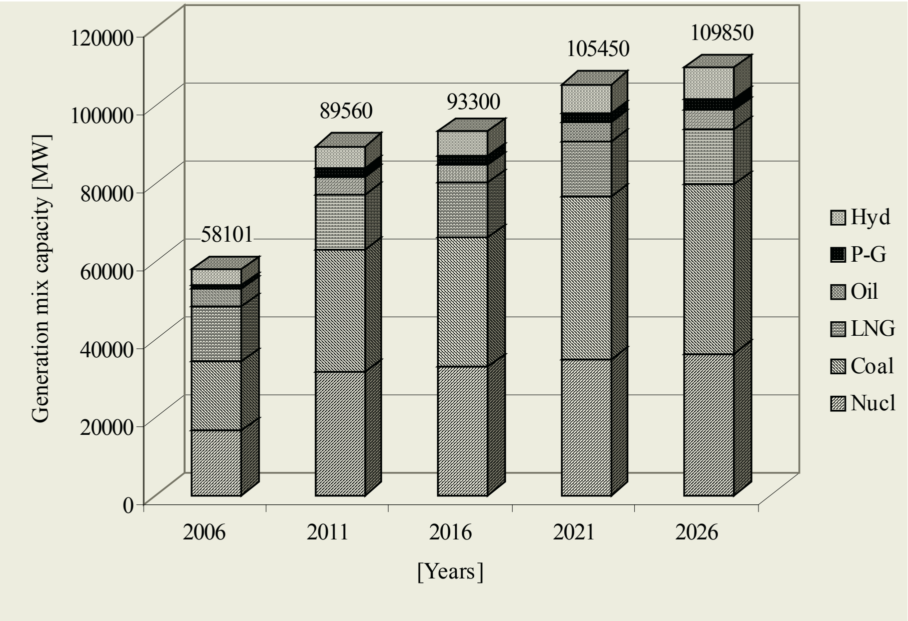
Table 5. Best generation mix in the three cases. [%]

Gen. type	Conventional method					Proposed method by fuzzy set theory							
						Mix without APC				APC reinforcement mix			
	2006	2011	2016	2021	2026	2011	2016	2021	2026	2011	2016	2021	2026
Nucl	28.8	35.2	35.5	32.9	33.1	35.1	35.6	33.4	33.8	37.1	39.3	36.4	36.6
Coal	30.1	34.9	35.2	39.6	39.5	35.1	35.5	38.3	38.5	31.4	30.2	28.4	28.9
LNG	24.6	16.0	15.3	13.6	13.0	16.1	15.3	13.6	13.0	17.2	15.5	18.5	17.8
Oil	7.4	4.8	4.6	4.6	4.4	4.9	4.6	5.7	5.4	5.1	4.6	5.1	4.9
P-G	1.7	2.6	2.5	2.2	2.7	2.3	2.1	1.9	1.8	2.4	3.6	4.5	4.6
Hyd	7.4	6.5	6.8	7.1	7.4	6.5	6.8	7.1	7.4	6.9	6.8	7.1	7.4

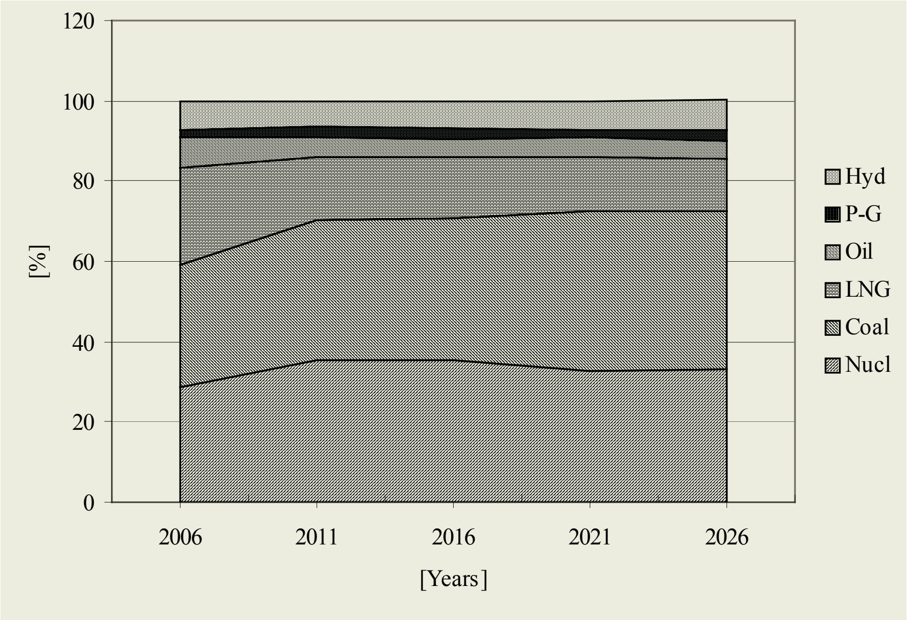
(where, APC means the air pollution(CO_2, SO_X and NO_X) constraints)

Fig. 3, Fig. 4 and Fig.5 shows total capacity and percent ratio results for conventional method and proposed method using fuzzy set theory.

The total cost evaluation given in Table 6 shows that there is much different total cost between the cases not considering air pollution constraints and the case considering air pollution constraints. As the air pollution constraints are considered, the mix of nuclear power plant is growing up and the total cost is increasing. It is interesting that the satisfaction level is decreasing.

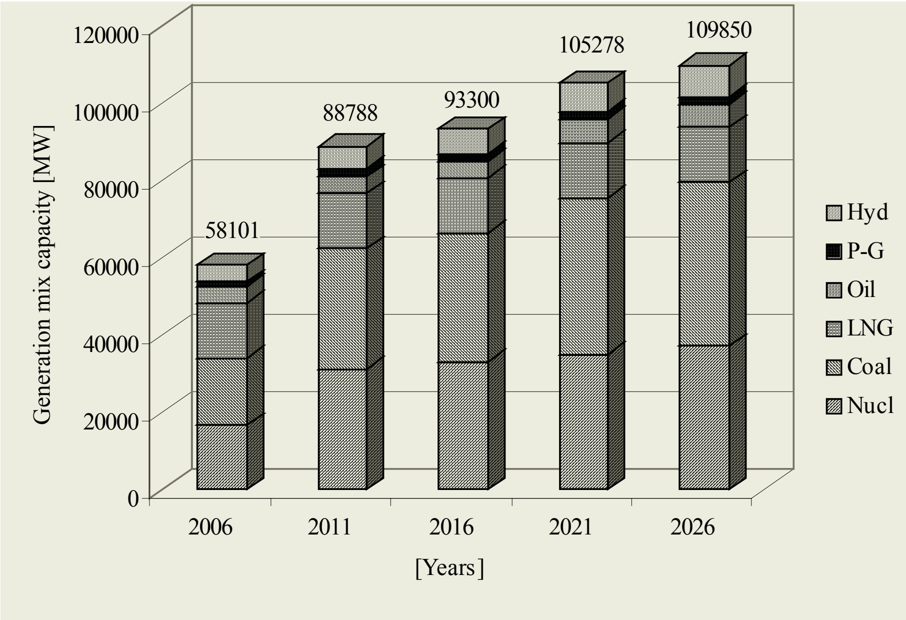


(a) Total capacity

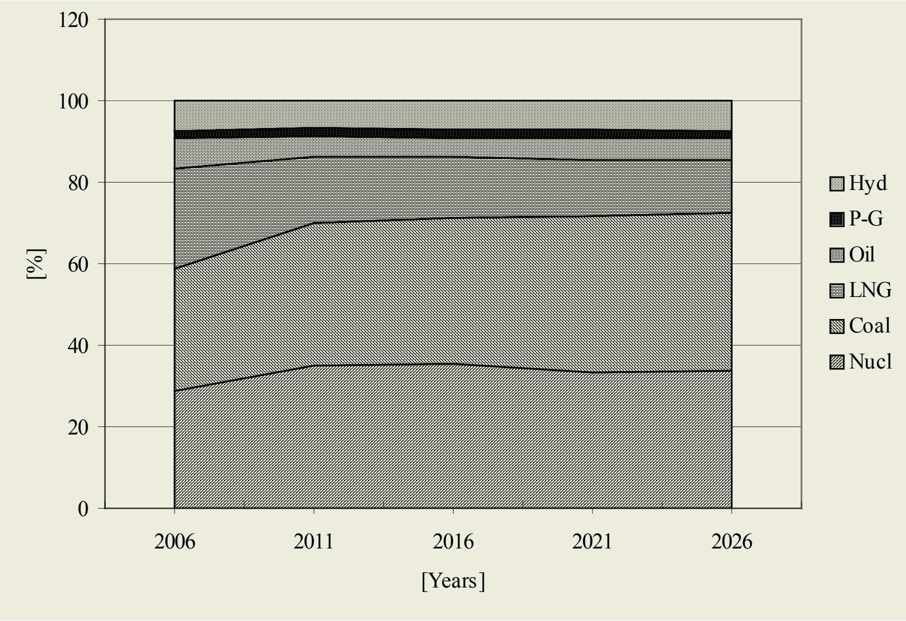


(b) Percent ratio

Fig. 3. Best generation mix by Conventional method

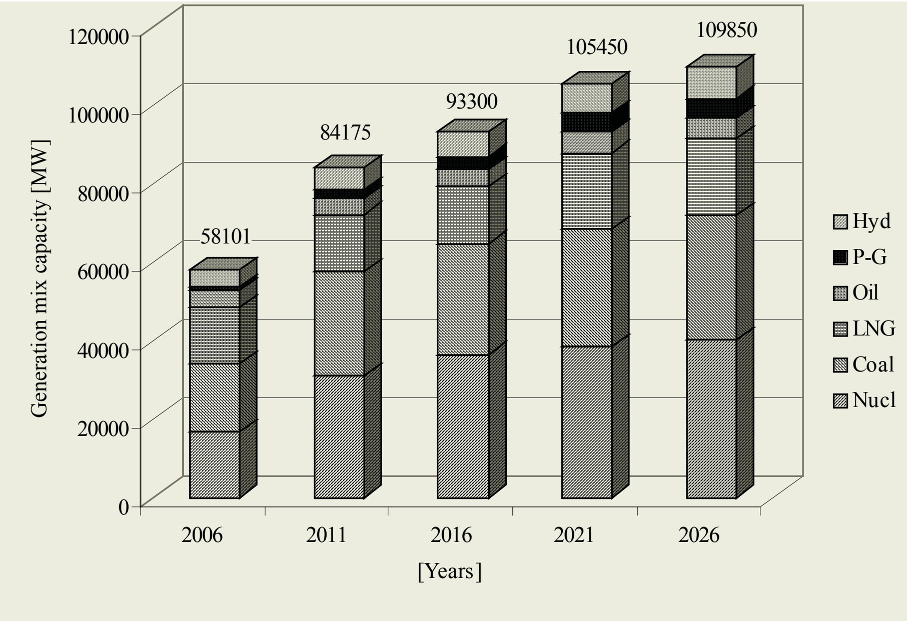


(a) Total capacity

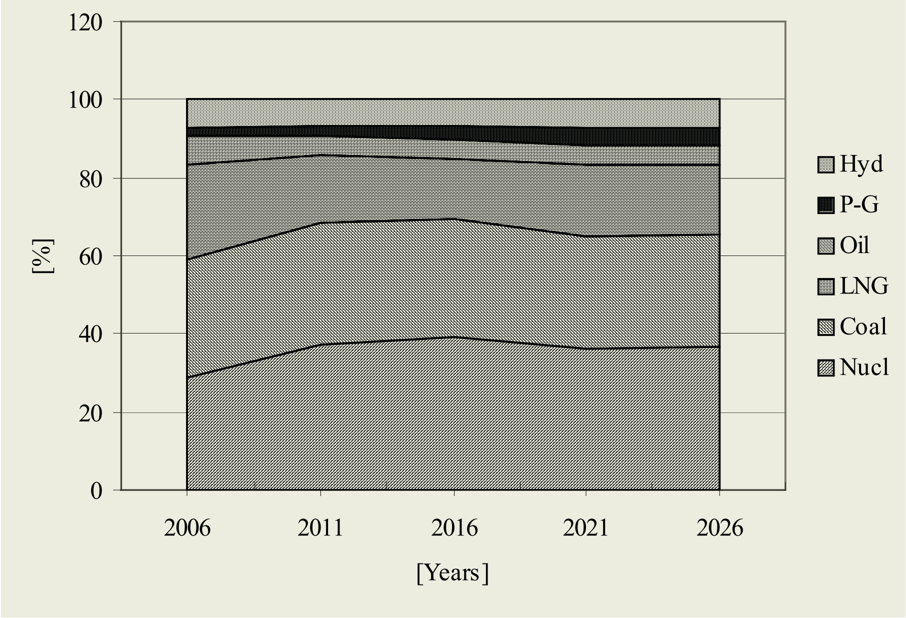


(b) Percent ratio

Fig. 4. Best generation mix by proposed fuzzy set theory method without air pollution constraint



(a) Total capacity



(b) Percent ratio

Fig. 5. Best generation mix by proposed fuzzy set theory method with air pollution constraint

Table 6. Total cost evaluation of best generation mix in the three cases. [Billion Won]

	Construction Cost	Operation Cost	Total Cost	Satisfaction Level
Conventional method	11,432.60	8,016.98	19,449.58	-
Mix without APC	11,602.87	8,211.28	19,814.14	0.96
Mix with APC	11,439.83	8,584.04	20,023.88	0.80

5 Conclusions

This paper proposes a new approach using the fuzzy linear programming for the long-term generation mix with air pollution constraints. The proposed approach is able to consider multi-criteria under uncertain circumstances associated with budget for construction and operation of generators, forecasted load, air pollution and energy resources by fuzzy set theory. A characteristic feature of the presented approach is that fuzziness in fuel and construction cost, load growth, reliability, air pollution can be taken into account by using the fuzzy linear programming. The multi-criteria can easily handled by not weighting factor method but the fuzzy theory. Additionally, case study shows that the realistic generation mix considering the air pollution constraints can be obtained.

Acknowledgement

This study was done by the Engineering Research Center (ERI) at Gyeongsang National University and the Electrical Power Reliability/Power Quality Research Center (EPRRC), which is one of the Electrical Industry Research Centers (EIRC) in Korea. The support of the Electrical Industry Research Center (EIRC) of the Ministry of Commerce, Industry and Energy (MOCIE) of Korea is acknowledged.

References

1. Hisham Khatib, *Economic Evaluation of Projects in the Electricity Supply Industry*, IEE Power & Energy Series 44., MPG Books Limited, Bodmin. Cornwall, 2003.
2. M. Ilic et al., *Power systems restructuring: Engineering and Economics*, Kluwer- Academic Pub., 1998.
3. Wang, J.R. McDonald: *Modern Power System Planning*. McGraw-Hill Book Company, 1994.
4. Young-Chang Kim, Byong-Hun Ahn: Multi-criteria Generation-Expansion Planning with Global Environmental Considerations, *IEEE Trans.,* Vol.40, May 1993, pp 154-161.
5. K. Yasuda, K. Nishiya, J. Hasegawa, R. Yokoyama: Optimal Generation Expansion Planning with Electric Energy Storage Systems: Industrial Electronics Society*, IECON '88. Proceedings*,Vol 3, October 1988, pp 550 – 555.
6. Kurt E. Yeager, *Power and Energy*, Spectrum, IEEE, Jan. 1996,Vol.33, Issue: 1, pp. 70-75.
7. Sasaki H., Kubokawa J., Watanabe M., Yokoyama R.; Tanabe R.: A solution of generation expansion problem by means of neutral network, *Neural Networks to Power Systems, Proceedings*, July 1991, pp 219 – 224.

8. Whei-Min L., Tung-Sheng Z., Ming-Tong T., Wen-Cha H.,: The generation expansion planning of the utility in a deregulated environment, Electric Utility Deregulation, Restructuring and Power Technologies, *Proceedings of the 2004 IEEE International Conference on.* Vol. 2, April 2004, pp 702 – 707.
9. Jinxiang Z., Mo-yuen C.: A review of emerging techniques on generation expansion planning, *IEEE Trans.*, Vol. 12, Nov. 1997, pp 1722 – 1728.
10. W.J.M. Kikert; *Fuzzy theories on decision-making*, Martinus Nihoff, 1978.
11. H.J. Zimmermann, *Fuzzy Set Theory and Its Applications*, Kluwer-Nijhohh Boston, pp.220-234, 1986.
12. Masatoshi Sakawa, *Fuzzy Sets and Interactive Multi-objective Optimization*, Plenum Press, New York, 1993.
13. J.S. Choi, D.H. Do, Development of a Method for the Generator Maintenance Scheduling using Fuzzy Integer Programming, *Journal of KFIS*, Vol. 7, No. 5, pp.77-85, 1997.
14. Hongsik Kim, Seungpil Moon, Jaeseok Choi, Soonyoung Lee, Daeho Do, and Madan M. Gupta. Generator Maintenance Scheduling Considering Air Pollution Based on the Fuzzy Theory, *IEEE International Fuzzy Systems Conference Proceedings*, August, Vol. III, pp. 1759~1764, 1999.
15. Kit Po W., Yin Wa W.: Combined genetic algorithm/simulated annealing/fuzzy set approach to short-term generation scheduling with take-or-pay fuel contract, *IEEE Trans.,* Vol.11, Feb. 1996, pp 128 – 136.

Ranking Functions, Perceptrons, and Associated Probabilities

Bernd-Jürgen Falkowski

University of Applied Sciences Stralsund
Zur Schwedenschanze 15
D-18435 Stralsund
Germany
Bernd.Falkowski@fh-stralsund.de

Abstract. The paper is motivated by a ranking problem arising e.g. in financial institutions. This ranking problem is reduced to a system of inequalities that may be solved by applying the perceptron learning theorem. Under certain additional assumptions the associated probabilities are derived by exploiting Bayes' Theorem. It is shown that from these a posteriori probabilities the original classifier may be recovered. On the other hand, assuming that perfect classification is possible, a maximum likelihood solution is derived from the classifier. Some experimental results are given.

1 Introduction

Ever since the Basel II central banks capital accord of the G10-states, cf. e.g. [1], the individual objective rating of the creditworthiness of customers has become an important problem. To this end so-called scoring systems, cf. e.g. [10], [5] have been used for quite some time. Generally these systems are implemented as linear discriminants where customer characteristics such as income, property assets, liabilities and the likes are assigned points or grades and then a weighted average is computed, where a customer is judged "good" or "bad" according to whether the average exceeds a cutoff point or not, cf. e.g. [13],[15]. In an extreme case the attributes are just binary ones where 0 respectively 1 signifies that the property does not hold respectively holds. The weights can then either be computed using classical statistical methods or more recently with the help of artificial neural networks provided that suitable bank records are available for training.

However, the use of only two classes for the classification of customers presents certain problems. On the one hand the event of a credit default is not precisely defined, cf. [1], p. 92, so that banking records would almost certainly need at least one more class (e.g. "doubtful (?) customers"). On the other hand the associated default probabilities (derived as a maximum likelihood solution in the case of two classes) are unfortunately not uniquely defined in general. This indicates that a finer distinction among customers could be useful. Indeed, after computation of these probabilities (again based on two classes) banks usually divide customers into a larger number of classes. This, of course, seems rather counter-intuitive, since surely the division should (and could) be based on experience and be effected before probabilities are

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1143–1150, 2005.

computed. Hence in this paper it is assumed that training data are available, where banking customers are divided into mutually disjoint risk classes $C_1, C_2, \ldots, C_n$. Here class C_i is preferred to C_j if $i<j$. It is then shown how this preference relation may be learned by a generalized perceptron. Note that the use of several classes has been investigated beforehand, see e.g. [2], p. 237. Moreover, the use of ranking functions has been recognized in an information retrieval context, cf, e.g. [18]. However, at least in the banking business, ranking functions, as described in section 2 below, see also [6], [16], apparently have not been used before. This is all the more surprising since this method allows the classification of customers according to creditworthiness before computing default probabilities instead of the other way round. At least from a theoretical point of view this seems much more satisfactory since that way the easier classification problem, cf. [17], is tackled before the more difficult problem of computing the probabilities. In addition the probabilities can be more precisely determined this way as experimental results given below show.

2 Reduction and Solution of the Ranking Problem

Suitable anonymous training data from a large German bank were available. In abstract terms then p vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ from $\Re^n$ (think of these as having grades assigned to individual customer characteristics as their entries) together with their risk classification (i.e. their risk class C_l for $1 \le l \le k$, where the risk classes were assumed to constitute a partition of pattern space) were given. Hence implicitly a preference relation (partial order) "$\rangle$" in pattern space was determined for these vectors by

$$\mathbf{x}_i \rangle \mathbf{x}_j \quad \text{if} \quad \mathbf{x}_i \in C_i \quad \text{and} \quad \mathbf{x}_j \in C_j \quad \text{where } i < j.$$

It was then required to find a map $m_{\mathbf{w}}: \Re^n \to \Re$ that preserves this preference relation, where the index $\mathbf{w}$ of course denotes a weight vector. More precisely one must have

$$\mathbf{x}_i \rangle \mathbf{x}_j \quad \Rightarrow m_{\mathbf{w}}(\mathbf{x}_i) > m_{\mathbf{w}}(\mathbf{x}_j)$$

If one now specializes by setting $m_{\mathbf{w}}(\mathbf{x}) := <\varphi(\mathbf{x}), \mathbf{w}>$, denoting the scalar product by $<.,.>$ and an embedding of $\mathbf{x}$ in a generally higher (m-) dimensional feature space by φ, then the problem reduces to finding a weight vector $\mathbf{w}$ and constants ("cut-offs") $c_1 > c_2 > \ldots > c_{k-1}$ such that

$$\begin{array}{lll} \mathbf{x} \in C_1 & \text{if } <\varphi(\mathbf{x}), \mathbf{w}> > c_1 & \\ \mathbf{x} \in C_l & \text{if } c_{l-1} \ge <\varphi(\mathbf{x}), \mathbf{w}> > c_l & \text{for} \quad l = 2, 3, \ldots, k-1 \\ \mathbf{x} \in C_k & \text{if } c_{k-1} \ge <\varphi(\mathbf{x}), \mathbf{w}>. & \end{array}$$

The problem may then be reduced further to a standard problem, whose solution may be obtained by various known algorithms, cf. e.g. [11],[7]:

Let $\mathbf{e}_i$ denote the i-th unit vector in $\Re^{k-1}$ considered as row vector and construct a matrix $\mathbf{B}$ of dimension $(m_1+2m_2+k-2)\times(m+k-1)$, where $m_1 := |C_1 \cup C_k|$ (here $|S|$ denotes the cardinality of set S) and $m_2 := |C_2 \cup C_3 \ldots \cup C_{k-1}|$, as follows:

$\mathbf{B} := \begin{bmatrix} \mathbf{R} \\ \mathbf{D} \end{bmatrix}$, dimension $R = (k-2) \times (m+k-1)$, and the i-th row of $\mathbf{R}$ is given by the row vector $(0, \ldots, 0, \mathbf{e}_i - \mathbf{e}_{i+1})$ with m leading zeros. Moreover $\mathbf{D}$ is described by:

For every vector $\mathbf{x}$ in C_1 respectively C_k $\mathbf{D}$ contains a row vector $(\boldsymbol{\varphi}(\mathbf{x}), -\mathbf{e}_1)$ respectively $(-\boldsymbol{\varphi}(\mathbf{x}), \mathbf{e}_{k-1})$, whilst for every vector $\mathbf{x}$ in C_l with $1 < l < k$ it contains the vectors $(\boldsymbol{\varphi}(\mathbf{x}), -\mathbf{e}_l)$ and $(-\boldsymbol{\varphi}(\mathbf{x}), \mathbf{e}_{l-1})$. The reduction of the problem to a system of inequalities is then proved by the following lemma.

Lemma 1: A weight vector $\mathbf{w}$ and constants $c_1 > c_2 > \ldots > c_{k-1}$ solving the ranking problem may (if they exist) be obtained by solving the standard system of inequalities $\mathbf{Bv} > \mathbf{0}$ where $\mathbf{v} := (\mathbf{w}, c_1, c_2, \ldots, c_{k-1})^T$.

Proof: Abbreviating the row vector $(\mathbf{w}, c_1, c_2, \ldots, c_{k-1})$ by $(\mathbf{w}, \mathbf{c})$ one computes:

(i) $<(0, \ldots, 0, \mathbf{e}_i - \mathbf{e}_{i+1}), (\mathbf{w}, c_1, c_2, \ldots, c_{k-1})> > 0$ for $i = 1, 2, ..., k-2$ implies that for the same i-values $c_i > c_{i-1}$ thus ensuring that the sequence (c_i) is monotonically decreasing.

(ii) If $\mathbf{x} \in C_1$ respectively C_k, then it follows from $<(\boldsymbol{\varphi}(\mathbf{x}), -\mathbf{e}_1), (\mathbf{w}, \mathbf{c}) > > 0$ respectively $<(-\boldsymbol{\varphi}(\mathbf{x}), \mathbf{e}_{k-1}), (\mathbf{w}, \mathbf{c}) > > 0$ that even $<\boldsymbol{\varphi}(\mathbf{x}), \mathbf{w}> > c_1$ respectively $c_{k-1} > <\boldsymbol{\varphi}(\mathbf{x}), \mathbf{w}>$ holds.

Moreover, if $\mathbf{x} \in C_l$ for $1 < l < k$, then for the same l-values $<(\boldsymbol{\varphi}(\mathbf{x}), -\mathbf{e}_l), (\mathbf{w}, \mathbf{c}) > > 0$ respectively $<(-\boldsymbol{\varphi}(\mathbf{x}), \mathbf{e}_{l-1}), (\mathbf{w}, \mathbf{c}) > > 0$ imply that even $c_{l-1} > <\boldsymbol{\varphi}(\mathbf{x}), \mathbf{w}> > c_l$ holds. Hence the lemma is proved. □

Of course, it must be admitted that the existence of a suitable weight vector $\mathbf{v}$ is by no means guaranteed. However, at least in theory, the map $\boldsymbol{\varphi}$ may be chosen such that the capacity of a suitable separating hyperplane is large enough for a solution to exist with high probability.

In order to see this note that the number of different monomials of degree i in $x_1, x_2, \ldots, x_n$ is given by B(n+i-1,i), where B denotes the binomial coefficient, cf. e.g. [12], p. 488. Hence an easy induction proof shows that the number of different monomials of degree less than or equal to i in $x_1, x_2, \ldots, x_n$ is given by B(n+i,i). Thus $\boldsymbol{\varphi}$ may, for example, be defined by

$$\boldsymbol{\varphi}\,(\mathbf{x}) := (1, x_1, x_2, \ldots, x_n, x_1^2, x_1x_2, \ldots, x_n^i)$$

as a map from $\mathfrak{R}^n$ to $\mathfrak{R}^{B(n+i,i)}$ and hence the separating capacity of the corresponding hyperplanes may be increased with i, for details see e.g. [3]. Thus, at least in theory, a suitable weight vector can be computed with high probability using for example the perceptron learning theorem (PLT), cf. e.g. [14], [11], if the map $\boldsymbol{\varphi}$ is chosen judiciously.

The price one has to pay for this increased separating capacity consists of larger computation times, that may, however, be acceptable if one uses the kernel version of the PLT, cf. [8], [16]. Perhaps, more importantly, a loss of generalization capabilities due to a higher VC-dimension of the separating hyperplanes, cf. e.g. [17], must be mentioned.

3 Associated Probability Distributions

In the case of just two preference classes a generalized linear classifier of the form described in section 2 may be arrived at, if the a priori class conditional probabilities (densities) are assumed to be members of the exponential distribution by applying Bayes' Theorem and moreover it is easily seen that the output of the associated perceptron may be interpreted as a probability if the logistic activation function is used, cf. e.g. [2], p.233. It seems natural to ask, if similar considerations apply here. Astonishingly enough, with some more or less obvious modifications this is indeed the case.

Define $D_l := \bigcup_{i=0}^{k-l} C_{k-i}$ for l = k, k-1, …, 2 and assume the probability (density) of **x** conditioned on D_l respectively D_l^C (the set-theoretic complement of D_l) to be given by members of the exponential distribution as follows:

$$p(\mathbf{x}|D_l) := \beta_l * \exp[B(\boldsymbol{\theta}_{l1}) + < \boldsymbol{\psi}(\mathbf{x}), \boldsymbol{\pi}(\boldsymbol{\theta}_{l1}) > + h_l(\mathbf{x})]$$
$$p(\mathbf{x}| D_l^C) := \beta_l * \exp[B(\boldsymbol{\theta}_{l2}) + < \boldsymbol{\psi}(\mathbf{x}), \boldsymbol{\pi}(\boldsymbol{\theta}_{l2}) > + h_l(\mathbf{x})]$$

where $\boldsymbol{\psi}(\mathbf{x}) := (\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \ldots, \psi_m(\mathbf{x}))$, $\boldsymbol{\pi}(\boldsymbol{\theta}_{l*}) := (\pi_1(\boldsymbol{\theta}_{l*}), \pi_2(\boldsymbol{\theta}_{l*}), \ldots, \pi_m(\boldsymbol{\theta}_{l*}))$, the β_l are normalizing constants, h_l, ψ_i are fixed functions from $\Re^n$ to $\Re$ and B, π_i, are fixed functions from some parameter space Θ to $\Re$, cf. e.g. [4], p.266.

In addition it is assumed that

$\boldsymbol{\pi}(\boldsymbol{\theta}_{l2}) - \boldsymbol{\pi}(\boldsymbol{\theta}_{l1})$ is a constant vector (independent of l) **w**, and that
$c_{l-1} := B(\boldsymbol{\theta}_{l1}) - B(\boldsymbol{\theta}_{l2}) - \ln[P(D_l^C)/P(D_l)$ is monotonically decreasing with l.

Here "ln" denotes the natural logarithm and $P(D_l^C)$ respectively $P(D_l)$ stand for the absolute probabilities of D_l^C respectively D_l. Note that notation has intentionally been abused here since **w** and c_l occurred in section 2 already.

With these definitions one easily obtains the following theorem.

Theorem 1:

$$P(C_1|\mathbf{x}) := g(<\boldsymbol{\psi}(\mathbf{x}), \mathbf{w}> - c_1)$$
$$P(C_l|\mathbf{x}) := g(<\boldsymbol{\psi}(\mathbf{x}), \mathbf{w}> - c_l) - g(<\boldsymbol{\psi}(\mathbf{x}), \mathbf{w}> - c_{l-1}) \text{ for } l = k-1, k-2, \ldots, 2.$$
$$P(C_k|\mathbf{x}) := 1 - g(<\boldsymbol{\psi}(\mathbf{x}), \mathbf{w}> - c_{k-1})$$

where $g(y):=[1+\exp(-y)]^{-1}$ denotes the logistic activation function.

Proof: This consists of an easy computation observing $D_k = C_k$, $D_2^C = C_1$, $P(C_{l-1}|x) = P(D_l^C|\mathbf{x}) - P(D_{l-1}^C|\mathbf{x})$ for l = k, k-1, ..., 3, and applying Bayes' Theorem. □

Exploiting the result of theorem 1 one can arrive at a classifier by stipulating that one should decide $\mathbf{x} \in C_i$ provided that $P(C_i|\mathbf{x}) > ½$ holds for i = 1, 2, …, k. This then leads to the following somewhat surprising result.

Corollary to theorem 1: Suppose that the c_i as described above are sufficiently far apart in a sense to be made precise below. Then $P(C_i|\mathbf{x}) > ½$ is approximately equivalent to demanding

$<\boldsymbol{\psi}(\mathbf{x}), \mathbf{w}> > c_1$	for	i = 1
$c_{i-1} \geq <\boldsymbol{\psi}(\mathbf{x}), \mathbf{w}> > c_i$	for	i = 2, 3, …, k-1
$c_{k-1} \geq <\boldsymbol{\psi}(\mathbf{x}), \mathbf{w}>$	for	i = k.

Proof: For the cases i = 1, k this is obvious. For the other cases clearly $P(C_i|\mathbf{x}) > ½$ implies $c_{i-1} \geq <\boldsymbol{\psi}(\mathbf{x}), \mathbf{w}> > c_i$, by considering theorem 1, recalling that the c_i are assumed to be monotonically decreasing and that g(0) = ½.

Next assume that c_i and c_{i-1} are far enough apart to make $g(<\boldsymbol{\psi}(\mathbf{x}), \mathbf{w}> - c_{i-1})$ negligibly small for $<\boldsymbol{\psi}(\mathbf{x}), \mathbf{w}>$ close to c_i. Hence, close to c_i, $<\boldsymbol{\psi}(\mathbf{x}), \mathbf{w}> > c_i$ must hold approximately. However, $f(y) := g(y - c_i) - g(y - c_{i-1})$ is at first monotonically increasing and then monotonically decreasing in the interval (c_i, c_{i-1}) as is easily seen by observing that its second derivative given by $f''(y) = g'(y - c_i)*[1 - 2g(y-c_i)] - g'(y - c_{i-1})*[1 -2g(y - c_{i-1})]$ is strictly negative in this interval and its first derivative at c_i satisfies $f'(c_i) \approx 1/4 > 0$. Thus for values of $<\psi(\mathbf{x}), \mathbf{w}>$ close to c_{i-1} ,$P(C_i|\mathbf{x}) > 1/2$ still holds provided that c_i is small enough (or c_i and c_{i-1} are far enough apart) to make $1- g(<\psi(\mathbf{x}), \mathbf{w}> - c_i)$ negligibly small. □

So by making suitable assumptions concerning the class conditional probabilities (densities) and applying Bayes' Theorem one arrives at a classifier as described in section 2. Note in this context, that if C_k is the default class, then the other classes are seen to be ordered by decreasing default probability, cf. theorem 1, which seems reasonable indeed. Of course, it is rather more interesting that as a consequence of these considerations the output of a generalized perceptron may be interpreted as probability if one uses a logistic activation function and makes the identifications indicated by the abuse of notation in theorem 1, where of course the function ψ appearing there corresponds to the function φ in section 2. This is all the more the case since a direct application of Bayes' Theorem appears difficult due to the hard to come by absolute probabilities.

4 Maximum Likelihood Solution

Following [2] we consider a network with one output y_l for each class C_l and target data $t_{ni} := \delta_{il}$ for the n-th pattern from class C_l. The probability of observing the target values given an input vector $\mathbf{x}_n$ is just $P(C_l|\mathbf{x}_n) = y_{nl}$. The value of the conditional distribution for this pattern is therefore given by

$$P(\mathbf{t}_n|\mathbf{x}_n) = \prod_{i=1}^{k} y_{ni}^{t_{ni}} .$$

Forming the likelihood function and taking the negative logarithm gives as error function

$$E = -\sum_{n} \sum_{i=1}^{k} t_{ni} \ln y_{ni} .$$

The absolute minimum of this error function occurs when $y_{ni} = t_{ni}$ for all values of i and n. It follows from theorem 1, that if a classifier as described in section 2 is computed first using e.g. the perceptron learning theorem then its output may be interpreted as probability by setting

$$P(C_1|\mathbf{x}) := g(<\varphi\ (\mathbf{x}), \mathbf{w}> - c_1)$$
$$P(C_l|\mathbf{x}) := g(<\varphi\ (\mathbf{x}), \mathbf{w}> - c_l) - g(<\varphi\ (\mathbf{x}), \mathbf{w}> - c_{l-1}) \text{ for } l = k-1, k-2, \ldots, 2.$$
$$P(C_k|\mathbf{x}) := 1 - g(<\varphi\ (\mathbf{x}), \mathbf{w}> - c_{k-1})$$

Actually the parameters $\mathbf{w}$ and $\mathbf{c}$ can only be determined up to a constant multiple, say α (>0). Hence the probabilities should have been given as

$$P(C_1|\mathbf{x}) := g(\alpha(<\varphi\ (\mathbf{x}), \mathbf{w}> - c_1))$$
$$P(C_l|\mathbf{x}) := g(\alpha(<\varphi\ (\mathbf{x}), \mathbf{w}> - c_l)) - g(\alpha(<\varphi\ (\mathbf{x}), \mathbf{w}> - c_{l-1})) \text{ for } l = k-1, k-2, \ldots, 2.$$
$$P(C_k|\mathbf{x}) := 1 - g(\alpha(<\varphi\ (\mathbf{x}), \mathbf{w}> - c_{k-1})).$$

Thus the following nice result is obtained.

Theorem 2: If the classifier computed according to section 2 allows perfect classification, then the associated probabilities

$$P(C_1|\mathbf{x}) := g(\alpha(<\varphi(\mathbf{x}), \mathbf{w}> - c_1))$$
$$P(C_l|\mathbf{x}) := g(\alpha(<\varphi(\mathbf{x}), \mathbf{w}> - c_l)) - g(\alpha(<\varphi(\mathbf{x}), \mathbf{w}> - c_{l-1})) \text{ for } l = k-1, k-2, \ldots, 2.$$
$$P(C_k|\mathbf{x}) := 1 - g(\alpha(<\varphi(\mathbf{x}), \mathbf{w}> - c_{k-1}))$$

constitute a maximum likelihood solution, if α is chosen sufficiently large.

Proof: Clearly $\mathbf{x} \in C_i \Rightarrow \lim P(C_i|\mathbf{x}) = 1$ as $\alpha \to \infty$. □

Remark: If perfect classification is not possible, then it seems reasonable to compute a suitable weight vector **w** and a cut-off vector **c** by a version of the so-called pocket algorithm, see e.g. [9], [7]. This algorithm works probabilistically and selects weights and a cut-off such that the number of errors is approximately minimized given a sufficiently long running time. It then remains to select the parameter α so as to maximize the likelihood function. It must be admitted that this procedure has not been entirely justified. However, as long as the error rate remains small one can reasonably hope that the perfect classification scenario gives a good indication of the true situation.

5 Experimental Results

Of course, in view of the possibly somewhat restrictive assumptions needed above it seemed desirable to conduct some experiments. These experiments were carried out with 361 data sets provided by a German financial institution. The customers had been divided into 4 preference classes and the method by which the classes had been obtained was not disclosed. Each customer was characterized by 8 attributes where each attribute had been assigned a grade (from 1 to 5, where 1 is the best grade) based on evaluations by internal computer programs (again the details of this evaluation were not disclosed). The experiments were conducted using a fault tolerant version of perceptron learning, cf. [7], on a standard laptop (3.2 GHz clock, 2 GB RAM). The CPU times required were considered negligible (less than 15 minutes in all cases), since only standard software (Excel and VBA) was employed for the implementation of the algorithm. For the tests 4 different initializations for the pocket algorithm were chosen since in previous tests with different data sets this had led to significantly differing weights for the classifier in the 2-class situation. Somewhat surprisingly in all cases a set of weights (and cut-offs) was obtained such that only 1 data set was erroneously classified. The sets of weights (including the cut-offs) were compared by computing the cosines of the angles between the vectors. The values obtained for the resulting 6 pairs were given by

Table 1. Cosines of angles between weight vectors

$\cos(v_1,v_2)$	$\cos(v_1,v_3)$	$\cos(v_1,v_4)$	$\cos(v_2,v_3)$	$\cos(v_2,v_4)$	$\cos(v_3,v_4)$
0,9999836	0,9999854	0,9999721	0,9999900	0,9999867	0,9999688

Here the cosine between v_i and v_j has, of course, been denoted by $\cos(v_i,v_j)$.

6 Discussion

Practical reasons have been given, why on the occasion of evaluating banking customers for creditworthiness more than 2 preference classes should be employed and why in this situation a ranking of customers is to be preferred to a mere classification. The resulting problem has been reduced to a standard problem and the associated probabilities have been computed. In addition it has been shown that under certain assumptions concerning the a priori probabilities the original ranking function can be recovered from the a posteriori probabilities obtained by an application of Bayes' Theorem. Moreover it is shown that from the classifier may, in the case of perfect separation, a maximum likelihood solution be obtained by choosing a smoothing parameter large enough. Finally experimental results obtained using genuine data indicate that the suggested procedure is indeed feasible. Although this evidence only refers to the scenario where almost perfect separation is possible one is inclined to think that, provided the error rate is not too large, similar results might hold for non-perfect separation. In conclusion then it seems necessary to conduct further practical experiments (if suitable data sets can be obtained) including validation tests with unseen data and in particular to investigate the accuracy of the probabilities computed from the ranking function. Indeed, the author is presently conducting negotiations with several large German banks who have expressed an interest in testing the proposed method.

References

1. Banking Committee on Banking Supervision: International Convergence of Capital Measurements and Capital Standards, A Revised Framework, Bank for International Settlements, http://www.bis.org/publ/bcbsca.htm (June 2004)
2. Bishop,C.M.: Neural Networks for Pattern Recognition. Oxford University Press, (1998)
3. Cover, T.M.: Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. IEEE Trans. on Electronic Computers, Vol. 14, (1965)
4. Devroye, L.; Györfi, L.; Lugosi, G. : A Probabilistic Theory of Pattern Recognition. Springer-Verlag, (1996)
5. Episcopos, A.; Pericli, A.; Hu, J.: Commercial Mortgage Default: A Comparison of the Logistic Model with Artificial Neural Networks. Proceedings of the Third International Conference on Neural Networks in the Capital Markets, London, England, (1995)
6. Falkowski, B.-J.: Lernender Klassifikator, Offenlegungsschrift DE 101 14874 A1, Deutsches Patent- und Markenamt, München, (Learning Classifier, Patent Number DE 101 14874 A1, German Patent Office, Munich) (2002)
7. Falkowski, B.-J.: Assessing Credit Risk Using a Cost Function. In: Proceedings of the Intl. Conference on Fuzzy Information Processing, Vol. II, Beijing, Tsinghua University Press, Springer-Verlag, (2003)
8. Falkowski, B.-J.: Scoring Systems, Classifiers, Default Probabilities, and Kernel Methods. In: Proceedings of the 2004 Intl. Conference on Machine Learning and Applications (ICMLA'04), Eds. M. Kantardzic, O. Nasraoui, M. Milanova, IEEE Catalog Number: 04EX970, (2004)
9. Gallant, S.I.: Perceptron-based Learning Algorithms. IEEE Transactions on Neural Networks, Vol. I, No. 2, (1990)
10. Hand, D.J.; Henley, W.E.: Statistical Classification Methods in Consumer Credit Scoring: a Review. J.R. Statist. Soc. A, 160, Part 3, (1997)

11. Haykin, S.: Neural Networks,.2nd edition, Prentice Hall, (1999)
12. Knuth, D.E.: The Art of Computer Programming. Vol. 1, Fundamental Algorithms, 2nd Edition, (1973)
13. Müller, M.; Härdle, W.: Exploring Credit Data. In: Bol, G.; Nakhneizadeh, G.; Racher, S.T.; Ridder, T.; Vollmer, K.-H. (Eds.): Credit Risk-Measurement, Evaluation, and Management, Physica-Verlag, (2003)
14. Ripley, B.D.: Pattern Recognition and Neural Networks. Oxford University Press, (1998)
15. Shadbolt, J.; Taylor, J.G.(Eds.): Neural Networks and the Financial Markets. Springer-Verlag, (2002)
16. Shawe-Taylor, J.; Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, (2004)
17. Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons, (1998)
18. Wong, S.K.M.; Ziarko, W.; Wong, P.C.N.: Generalized Vector Space Model in Information Retrieval. Proceedings of the 8th ACM SIGIR Conference on Research and Development in Information Retrieval, USA, (1985)

Directed Mutation Operators – An Overview

Stefan Berlik and Bernd Reusch

University of Dortmund, Department of Computer Science,
D-44221 Dortmund, Germany
{Stefan.Berlik,Bernd.Reusch}@udo.edu

Abstract. Directed mutation has shown to improve the efficiency of evolutionary algorithms significantly for a broad spectrum of optimization problems. When the first mutation operators of this kind, however, suffered from the asymmetry parameter influencing the mutation strength, in the meantime there are several new directed mutation operators available which overcome this drawback. The aim of this paper is to give an overview of all different operators in one single place. Their characteristics will be presented and their advantages and disadvantages are discussed. At the end a comparison and a summary is provided.

1 Introduction

The capability of directed mutation has been shown for both, real world applications as well as for test problems [5]. Because of their increasing importance and the number of different directed mutation operators being available now, a comparison seems to be expedient.

The main idea of the directed mutation is to mutate with random numbers that lie preferably in the direction the optimum is expected. This implies expected values unequal to zero and thereby contrasts the classical mutation operators, but using this method the optimization strategy can adopt the most promising direction over the generations.

To do so a customizable skewed distribution is needed, whereof some will be presented here.

2 Directed Mutation Operators

To be able to do directed mutation one needs to generate skew distributed random numbers. Thereby expected values unequal to zero are introduced. This means that the mutation operator is not compliant to standard evolution strategy any longer postulating an expected value of zero, so mutating around the origin. Further it has to be ensured that the expected value is convergent, thus forcing the mutation operator to continue mutating near by the origin. It should be pointed out that convergence in the symmetrical case only is not enough. Convergence has rather to be guaranteed for all skewed cases. Mutation operators based on the Ξ-distribution proposed by Hildebrand [7] violate this demand [4].

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1151–1159, 2005.

Diverging expected values caused by increasing skewness parameters here can lead to wide jumps. Of even greater interest is the variance. It can be seen as a measure for the mutation strength and is a strategy parameter in most evolution strategies. Because of this it should not be modified by the skewness parameter. In the ideal case the variance is an invariant, independent of the skewness parameter. At least convergence is necessary and a small spread between minimal and maximal value is desired to limit the impact of the skewness on the mutation strength. Again, the Ξ-distribution violates this demand. To overcome this, several alternative directed mutation operators have been developed, all with convergent moments. The first one was the naive skew-normal mutation that is strongly related to the asymmetric mutation. A completely different approach follow the skew-normal mutation operators that will be presented afterwards.

3 Asymmetric Mutation

For his asymmetric mutation Hidebrand has chosen an additive approach [7]. The density function is defined in sections and made up of two parts, one function for the negative domain and another function for the positive domain, where one of these functions is always exactly the standard normal density. The other one is an expanded normal density. To transform the whole function into a density the integral is then normalized to one.

3.1 Probability Density Function

The complete definition of the density of the Ξ_c-distribution splits into four cases and is given below:

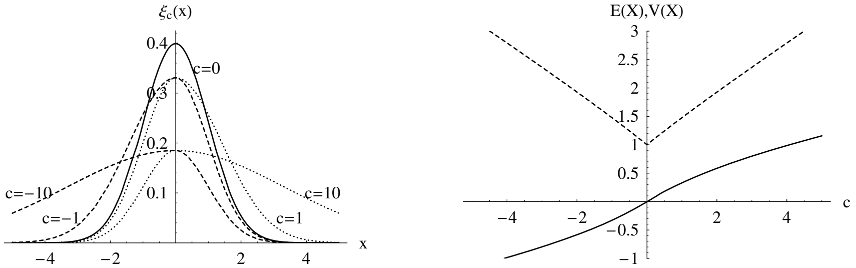
$$\xi_{c,\sigma}(x) = \begin{cases} \frac{\sqrt{2}}{\sqrt{\pi\sigma}\left(1+\sqrt{1-c}\right)}\, e^{-\frac{x^2}{2\sigma(1-c)}} & \text{if } c < 0, x < 0 \\ \frac{\sqrt{2}}{\sqrt{\pi\sigma}\left(1+\sqrt{1-c}\right)}\, e^{-\frac{x^2}{2\sigma}} & \text{if } c < 0, x \geq 0 \\ \frac{\sqrt{2}}{\sqrt{\pi\sigma}\left(1+\sqrt{1+c}\right)}\, e^{-\frac{x^2}{2\sigma}} & \text{if } c \geq 0, x < 0 \\ \frac{\sqrt{2}}{\sqrt{\pi\sigma}\left(1+\sqrt{1+c}\right)}\, e^{-\frac{x^2}{2\sigma(1+c)}} & \text{if } c \geq 0, x \geq 0. \end{cases} \tag{1}$$

For some values of c, ξ_c density functions are plotted in Fig. 1(a). Note the fat tails which lead to the diverging expectation and variance.

3.2 Moments

Given below are the formulae of the expected value and the variance of a Ξ_c distributed random variable X.

$$\mathrm{E}(X) = \sqrt{\frac{2}{\pi}} \frac{c}{1 + \sqrt{1 + |c|}} \tag{2}$$



(a) The density functions $\xi_{-10}, \xi_{-1}, \xi_0, \xi_1$, and ξ_{10}

(b) Expectation (solid) and variance (dashed) of a Ξ_c distributed random variable

Fig. 1. Asymmetric mutation

$$\mathrm{Var}(X) = 2 - \sqrt{1+|c|} + |c| - \frac{2c^2}{\pi\left(1+\sqrt{1+|c|}\right)^2} \tag{3}$$

As mentioned before, the expected value as well as the variance diverges (4). Their graphs are depicted in Fig. 1(b).

$$\lim_{c\to\pm\infty}\big(\mathrm{E}(Z)\big) = \mathrm{sign}(c)\,\infty \qquad \lim_{c\to\pm\infty}\big(\mathrm{Var}(Z)\big) = \infty. \tag{4}$$

3.3 Random Variate Generation

Generating Ξ_c distributed random numbers is demanding and cumbersome. It is done using the inverse function of the Ξ_c-distribution. Random numbers thus are created by multiplying uniform distributed random numbers with the inverse function. The inverse distribution is defined as

$$\overline{\Xi}_c(y) = \begin{cases} \sqrt{2(1-c)}\ \mathrm{inverf}\left(y\left(1+\frac{1}{\sqrt{1-c}}\right)-1\right) & \text{if } c<0, y<\frac{\sqrt{1-c}}{1+\sqrt{1-c}} \\ \sqrt{2}\ \mathrm{inverf}\left(y\left(1+\sqrt{1-c}\right)-\sqrt{1-c}\right) & \text{if } c<0, y\geq\frac{\sqrt{1-c}}{1+\sqrt{1-c}} \\ \sqrt{2}\ \mathrm{inverf}\left(y\left(1+\sqrt{1+c}\right)-1\right) & \text{if } c\geq 0, y<\frac{1}{1+\sqrt{1+c}} \\ \sqrt{2(1+c)}\ \mathrm{inverf}\left(y+\frac{y-1}{\sqrt{1+c}}\right) & \text{if } c\geq 0, y\geq\frac{1}{1+\sqrt{1+c}} \end{cases} \tag{5}$$

Note that there are two case differentiations, one calculation of the transcendent inverse error function, and several arithmetic operations necessary to generate a Ξ_c distributed random number.

4 Naive Skew-Normal Mutation

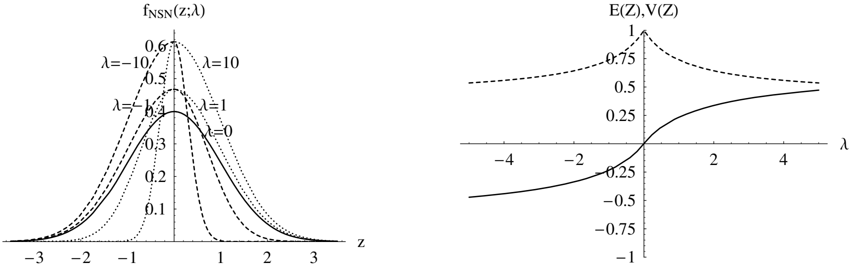
The naive skew-normal (NSN, hereafter) distribution is built in a similar manner to the Ξ_c-distribution [6]. The main difference lies in compressing on half of the normal density instead of expanding it. This avoids fat tails (see Fig. 2(a)) and guarantees convergent expectation and variance (9).

4.1 Probability Density Function

A random variable Z is said to be naive skew-normal with parameter λ , written $Z \sim NSN(\lambda)$, if its probability density function is

$$f_{NSN}(z;\lambda) = \begin{cases} \sqrt{\frac{2}{\pi}} \frac{\sqrt{1-\lambda}}{\left(1+\sqrt{1-\lambda}\right)} \, e^{-\frac{1}{2}z^2} & \text{if } \lambda \leq 0, z \leq 0 \\ \sqrt{\frac{2}{\pi}} \frac{\sqrt{1-\lambda}}{\left(1+\sqrt{1-\lambda}\right)} \, e^{-\frac{(1-\lambda)}{2}z^2} & \text{if } \lambda \leq 0, z > 0 \\ \sqrt{\frac{2}{\pi}} \frac{\sqrt{1+\lambda}}{\left(1+\sqrt{1+\lambda}\right)} \, e^{-\frac{(1+\lambda)}{2}z^2} & \text{if } \lambda > 0, z \leq 0 \\ \sqrt{\frac{2}{\pi}} \frac{\sqrt{1+\lambda}}{\left(1+\sqrt{1+\lambda}\right)} \, e^{-\frac{1}{2}z^2} & \text{if } \lambda > 0, z > 0. \end{cases} \tag{6}$$

Graphs for several degrees of skewness of the NSN density are shown in Fig. 2(a).



(a) The density functions NSN(-10), NSN(-1), NSN(0), NSN(1), and NSN(10)

(b) Expectation (solid) and variance (dashed) of a NSN(λ) distributed random variable

Fig. 2. Naive skew-normal mutation

4.2 Moments

The formulae for the expected value and the variance of a NSN distributed random variable Z take the following form:

$$\mathrm{E}(Z) = \sqrt{\frac{2}{\pi}} \frac{\lambda}{1+|\lambda|+\sqrt{1+|\lambda|}} \tag{7}$$

$$\mathrm{Var}(Z) = \frac{4\left(\sqrt{1+|\lambda|}-1\right)+|\lambda|\,(\pi-2)+\pi\left(2-\sqrt{1+|\lambda|}\right)}{\pi\,(1+|\lambda|)} \tag{8}$$

The limits are:

$$\lim_{\lambda\to\pm\infty}\big(\mathrm{E}(Z)\big) = \operatorname{sign}(\lambda)\sqrt{\frac{2}{\pi}} \qquad \lim_{\lambda\to\pm\infty}\big(\mathrm{Var}(Z)\big) = \frac{\pi-2}{\pi} \tag{9}$$

Their graphs are depicted in Fig. 2(b). One can see that the variance is convergent, but still spreads about 0.64. To make the variance an invariant, a linear transformation has to be applied to the NSN distributed random variable leading to the standardized NSN distribution.

4.3 Random Variate Generation

NSN distributed random variables can be generated using the method described in 3.3 with the appropriate inverse distribution given below.

$$\overline{F}_{NSN}(y;\lambda) = \begin{cases} \sqrt{2}\,\operatorname{inverf}\left(y\left(1+\frac{1}{\sqrt{1-\lambda}}\right)-1\right) & \text{if } \lambda<0, y<\frac{\sqrt{1-\lambda}}{1+\sqrt{1-\lambda}} \\ \frac{\sqrt{2}}{\sqrt{1-\lambda}}\,\operatorname{inverf}\left(y\left(1+\sqrt{1-\lambda}\right)-\sqrt{1-\lambda}\right) & \text{if } \lambda<0, y\geq\frac{\sqrt{1-\lambda}}{1+\sqrt{1-\lambda}} \\ \frac{\sqrt{2}}{\sqrt{1+\lambda}}\,\operatorname{inverf}\left(y\left(1+\sqrt{1+\lambda}\right)-1\right) & \text{if } \lambda\geq 0, y<\frac{1}{1+\sqrt{1+\lambda}} \\ \sqrt{2}\,\operatorname{inverf}\left(y\left(1+\frac{1}{\sqrt{1+\lambda}}\right)-\frac{1}{\sqrt{1+\lambda}}\right) & \text{if } \lambda\geq 0, y\geq\frac{1}{1+\sqrt{1+\lambda}} \end{cases} \tag{10}$$

5 Standardized Naive Skew-Normal Mutation

Obviously the NSN distribution can be transformed into a version with invariant variance. The standardization term that has to be applied is

$$\sigma_{Std}(\lambda) = \sqrt{\frac{\pi(1+|\lambda|)}{4\left(\sqrt{1+|\lambda|}-1\right)+|\lambda|\,(\pi-2)+\pi\left(2-\sqrt{1+|\lambda|}\right)}}\,. \tag{11}$$

For further details on the standardized naive skew-normal mutation see [6].

6 Skew-Normal Mutation

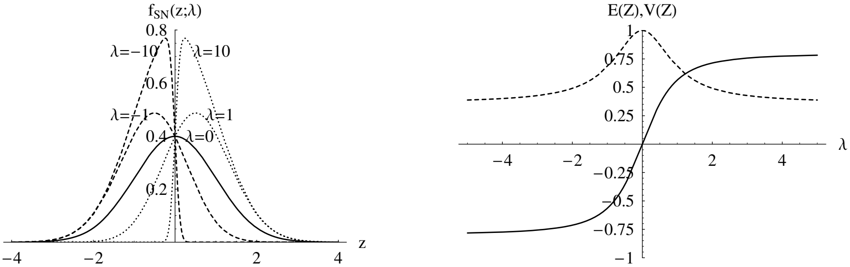
The class of distributions that is used to build the following directed mutation operator is called skew-normal (SN) distribution and was introduced by Azzalini [2]. A detailed presentation of the SN distribution, some extensions, and a small historical review are given by Arnold and Beaver [1].

6.1 Probability Density Function

The SN density function is defined by

$$f_{SN}(z;\lambda) = 2\varphi(z)\Phi(\lambda z) \tag{12}$$

where φ and Φ represents the probability density function and the cumulative distribution function of the standard normal density, respectively. λ is a real parameter that controls the skewness, where positive (negative) values indicate positive (negative) skewness. In the case $\lambda = 0$ the SN density gets back to the normal density (see Fig. 3(a)). With $Z \sim SN(\lambda)$ one denotes a random variable that has density (12).



(a) The density functions SN(-10), SN(-1), SN(0), SN(1), and SN(10)

(b) Expectation (solid) and variance (dashed) of a SN(λ) distributed random variable

Fig. 3. Skew-normal mutation

6.2 Moments

The first four moments are given by

$$\mathrm{E}(Z) = b\delta \tag{13}$$

$$\mathrm{Var}(Z) = 1 - (b\delta)^2 \tag{14}$$

$$\gamma_1(Z) = \frac{1}{2}(4-\pi)\,\mathrm{sign}(\lambda)\left(\frac{\big(\mathrm{E}(Z)\big)^2}{\mathrm{Var}(Z)}\right)^{3/2} \tag{15}$$

$$\gamma_2(Z) = 2(\pi-3)\left(\frac{\big(\mathrm{E}(Z)\big)^2}{\mathrm{Var}(Z)}\right)^{2} \tag{16}$$

where $b = \sqrt{2/\pi}$ and $\delta = \lambda/\sqrt{1+\lambda^2}$.

$\gamma_1(Z)$ and $\gamma_2(Z)$ denote the skewness and kurtosis. As desired, both expectation and variance converge. The limits are

$$\lim_{\lambda\to\pm\infty}\big(\mathrm{E}(Z)\big) = \mathrm{sign}(\lambda)\sqrt{\frac{2}{\pi}} \qquad \lim_{\lambda\to\pm\infty}\big(\mathrm{Var}(Z)\big) = 1 - \frac{2}{\pi}. \tag{17}$$

Their graphs are depicted in Fig. 3(b). One can see that the variance is convergent, but still spreads about 0.64. To make the variance invariant, a linear transformation has to be applied to the SN distributed random variable leading to the standardized SN distribution.

6.3 Random Variate Generation

Generation of SN distributed random numbers is simple and fast. A random variable Z with density (12) can be generated by an acceptance-rejection method [3]. Therefore sample Y and W from φ and Φ', respectively. Then Z is defined to be equal to Y or $-Y$, conditionally on the event $\{W \leq \lambda Y\}$:

$$Z = \begin{cases} Y & \text{if } W \leq \lambda Y \\ -Y & \text{if } W > \lambda Y \end{cases} \tag{18}$$

Thus simply two standard normal random variables are needed to generate one SN distributed random variable.

7 Standardized Skew-Normal Mutation

Using a linear transformation the SN distribution can be changed to a version where the skewness does not influence the variance any longer. The variance then becomes an invariant [5]. This is achieved using the transformed random variable sZ with

$$s = \frac{1}{\sqrt{V(Z)}} = \frac{1}{\sqrt{1-(b\delta)^2}} = \sqrt{\frac{\pi(1+\lambda^2)}{\pi+(\pi-2)\lambda^2}}. \tag{19}$$

7.1 Probability Density Function

The density of the standardized skew-normal distribution (SSN) is

$$f_{SSN}(z;\lambda) = \frac{2}{s}\varphi\left(\frac{z}{s}\right)\Phi\left(\frac{\lambda z}{s}\right). \tag{20}$$

For some values of λ, SSN density functions are plotted in Fig. 4(a). Note that due to the standardization the densities are flattened and widened.

7.2 Moments

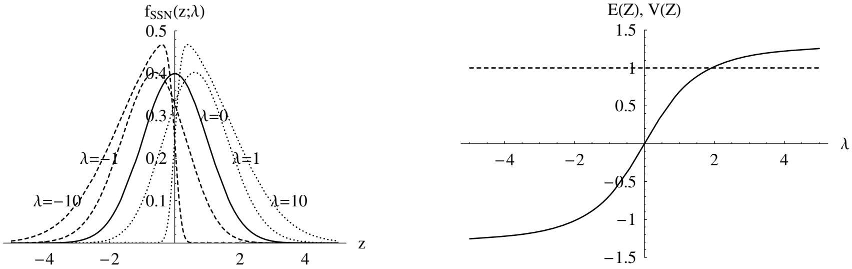
The first four Moments of the SSN distribution can be deduced from the moments of the SN distribution (13) – (16), leading to

$$\mathrm{E}(Z) = sb\delta \tag{21}$$

$$\mathrm{Var}(Z) = 1 \tag{22}$$

$$\gamma_1(Z) = \frac{1}{2}(4-\pi)\big(\mathrm{E}(Z)\big)^3 \tag{23}$$

$$\gamma_2(Z) = 2(\pi-3)\big(\mathrm{E}(Z)\big)^4. \tag{24}$$



(a) The density functions SSN(-10), SSN(-1), SSN(0), SSN(1), and SSN(10)

(b) Expectation (solid) and variance (dashed) of a SSN(λ) distributed random variable

Fig. 4. Standardized skew-normal mutation

The limits of the SSN class are

$$\lim_{\lambda \to \pm\infty} \big(\mathrm{E}(Z)\big) = \mathrm{sign}(\lambda)\sqrt{\frac{2}{\pi - 2}} \qquad \lim_{\lambda \to \pm\infty} \big(\mathrm{Var}(Z)\big) = 1. \tag{25}$$

Graphs of the first two moments are shown in Fig. 4(b).

7.3 Random Variate Generation

SSN random variables can be generated with the method described in 6.3.

8 Summary

The presented results are summarized in Table 1. One sees that SN and SSN perform considerably better than the other mutation operators. While all but the asymmetric mutation operator have convergent expectation and variance these two are the only which provide acceptable random variate generation procedures. Taking into account that during an optimization process a vast amount of random numbers has to be generated this issue is very important. This is also reflected in the point *Usefulness.* The asymmetric mutation is unusable because of its diverging moments, whereas the head start of the SN and SSN mutations originates from the random number generation.

9 Conclusions

With the directed mutation by means of the SN or the SSN distribution mutation operators are given that clearly outperform the others and also the classical mutation operators. These two are the only directed mutation operators with the density function given in closed form and simple and fast random number

Table 1. Comparison of the mutation operators

	Asymmetric	Naive SN	Std. naive SN	Skew-normal	Std. SN
Convergent expectation	-	+	+	+	+
Convergent variance	-	+	+	+	+
Invariant variance	-	-	+	-	+
Mathematical tractability	o	o	o	+	+
Given in closed form	-	-	-	+	+
Random variate generation	o	o	o	+	+
Usefulness	-	o	o	+	+

generators. Taking into account that the algorithm itself is quit fast (e.g. compared to the correlated mutation [8]) the use of the SN or SSN directed mutation might be quite beneficial for many problems. If an invariant variance is desired the SSN should be used, causing only a slight overhead compared to the SN mutation.

References

1. B. C. Arnold and R. J. Beaver. Skewed multivariate models related to hidden truncation and/or selective reporting. *Test. Sociedad de Estadística e Investigación Operativa*, 11(1):7–54, 2002.
2. A. Azzalini. A class of distributions which includes the normal ones. *Scand. J. Statist.*, 12:171–178, 1985.
3. A. Azzalini. Further results on a class of distributions which includes the normal ones. *Statistica*, 46:199–208, 1986.
4. S. Berlik. A polymorphical mutation operator for evolution strategies. In M. Wagenknecht and R. Hampel, editors, *Proc. of the Int. Conf. in Fuzzy Logic and Technology, EUSFLAT*, pages 502–505. European Society for Fuzzy Logic and Technology, EUSFLAT, 2003.
5. S. Berlik. Directed mutation by means of the skew-normal distribution. In *Proc. of the Int. Conf. on Computational Intelligence, FUZZY DAYS*, LNCS. Springer-Verlag Berlin Heidelberg, 2004.
6. S. Berlik. A directed mutation framework for evolutionary algorithms. In R. Matoušek and P. Ošmera, editors, *Proc. of the 10th Int. Conf. on Soft Computing, MENDEL*, pages 45–50, 2004.
7. L. Hildebrand. *Asymmetrische Evolutionsstrategien.* PhD thesis, Universität Dortmund, 2001.
8. H.-P. Schwefel. *Evolution and Optimum Seeking.* John Wiley & Sons, New York, 1995.

A General Fuzzy Min Max Neural Network with Compensatory Neuron Architecture

A.V. Nandedkar and P.K. Biswas

Indian Institute of Technology,
Kharagpur - 721302, India
{avn,pkb}@ece.iitkgp.ernet.in
http://www.iitkgp.ernet.in

Abstract. This paper proposes “A General Fuzzy Min-max neural network with Compensatory Neurons architecture”(GFMCN) for pattern classification and clustering. The network is capable of handling labeled and unlabeled data simultaneously, on-line. The concept of compensatory neurons is inspired from reflex system of the human brain. Fuzzy min-max neural network based architectures use fuzzy hyperbox sets to represent the data cluster or classes. An important stage in the training phase of these techniques is to manage the hyperbox overlaps and containments. In case of GFMCN, compensatory neurons are trained to handle the hyperbox overlap and containment. Inclusion of these neurons with a new learning approach has improved the performance significantly for labeled as well as unlabeled data. Moreover accuracy is almost independent of the maximum hyperbox size. The advantage of GFMCN is that it can learn data in a single pass (on-line). The performance of GFMCN is compared with “General Fuzzy Min-max neural network” proposed by Gabrys and Bargiela on several datasets.

1 Introduction

Many computer vision applications demand understanding and classifying data patterns for making certain decisions or actions. Hence pattern classification is an important step in the field of computer vision. The fundamental objective of pattern classification is to identify the underlying structure in the data. Fuzzy interpretation of patterns is very natural in the cases where precise partitions of data are not known. The seminal paper of Zadeh [1] elaborates the importance of fuzzy logic for pattern classification.

The merge of fuzzy logic and ART1 [9] neural network for pattern clustering has been reported in [10], [3]. Simpson proposed the fuzzy min-max neural network (FMNN) for classification [2] and clustering [3] using fuzzy hyperbox set concept, which is a powerful method to classify the linearly non-separable multidimensional data in a single pass. The collection of the hyperboxes belonging to the same class can classify linearly non-separable n-dimensional data very easily. The recursive GFMN algorithm [5] proposed by Gabrys and Bargiela is a merge of FMNN classification and FMNN clustering algorithms. The main advantage of GFMN is that the hybrid approach enabled the handling of labeled and unlabeled data simultaneously. The proposed GFMCN classifier is also based on the hyperbox fuzzy set concept. It

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1160–1167, 2005.

handles labeled and unlabeled data efficiently. GFMCN uses Compensatory Neurons (CN) [11] to overcome the hyperbox overlap and containment problems. CNs are inspired from the reflex system of human brain [4]. The paper also proposes a modified learning algorithm for *labeling* the *unlabeled* data while training.

The error analysis for FMNN due to contraction process in learning phase is given in [11]. This analysis is equally valid for GFMN algorithm as it also includes similar contraction process. Performance of FMNN and GFMN are highly dependant on the choice of expansion coefficient. There are many approaches reported in the literature to optimize the performance of FMNN classifier such as [6]-[8]. However, the optimization in these methods is achieved at the cost of recursive nature of the algorithms i.e. single pass through learning capability is lost.

The proposed GFMCN maintains the single pass though, on-line learning capability. This is achieved by removing the contraction process for the labeled hyperboxes. Instead of hyperbox contraction, compensatory neurons are added to the network dynamically during training. These compensatory neurons maintain the hyperbox dimensions and control the membership in the overlapped region.

The rest of the paper is organized as follows. Second section gives the details of GFMCN architecture. Section 3 provides the learning and recall algorithms for GFMCN. In section 4, a comparison of FMNN, GFMN and GFMCN in different modes (classifier, hybrid modes) of operating on real data sets is presented. The final section concludes with summary.

2 General Fuzzy Min-Max Classifier with Compensatory Neuron Architecture

Figure 1 shows the neural network architecture of General Fuzzy Min-max neural network with Compensatory Neuron (GFMCN). It consists of three sections: 1) The classifying neuron section,(CL) 2) The overlap compensating section (OC)and 3) The containment compensating (CC) section.

An n-dimensional normalized data a_{h1}-a_{hn} is fed to input nodes a_1-a_n. The neurons b_1-b_m are classifying neurons. The class node C_0 represents all ***unlabeled*** hyperboxes in the middle layer of CL section. Outputs of the CNs belonging to a class are collected at a class node C_i in the output layer. The nodes d_1-d_p are overlap compensation neurons (OCN) and e_1-e_q represent the containment compensation neurons (CCN). Outputs of OCN & CCN are collected at a class node C_i in respective compensation sections. Note that there is no compensation signal for class 0 i.e. for unlabeled hyperboxes.

The activation function of the **classifying neuron** b_j [5] is given by,

$$b_j(a_h,V_j,W_j) = \min_{i=1..n} (\min[(1-f(a_{hi}-W_{ji},\gamma)),(1-f(V_{ji}-a_{hi},\gamma))]) \tag{1}$$

where V, W: min-max point of the hyperbox b$_j$. γ: Fuzziness control parameter. $f(x,\gamma)$: is the two parameter ramp threshold function, n- dimension of data.

$$f(x,\gamma) = \begin{cases} 1 & if\ x\gamma > 1 \\ x\gamma & if\ 0 \le x\gamma \le 1 \\ 0 & if\ x\gamma < 0 \end{cases} \tag{2}$$

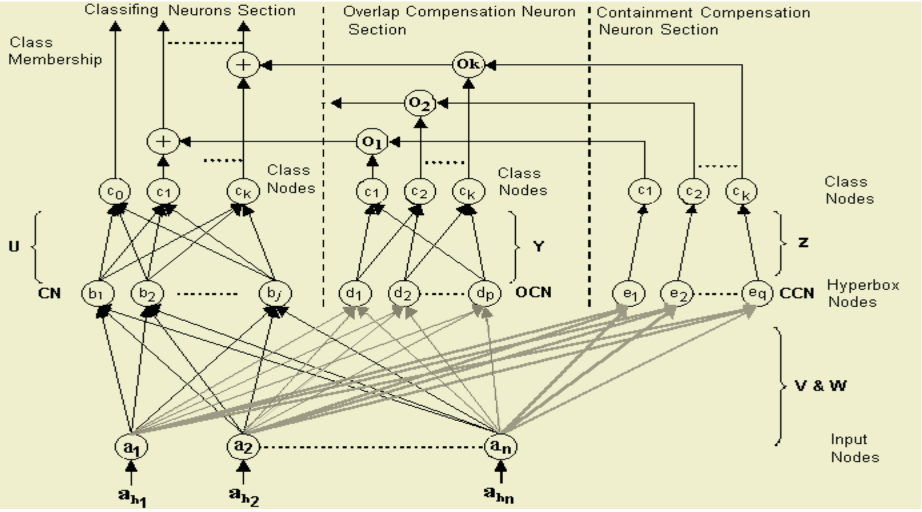


Fig. 1. G FMCN Architecture

The maximum hyperbox size is controlled by the expansion coefficient Θ,

$$n\Theta \geq \sum_{i=1}^{n}(max(w_{ji}, a_{hi}) - min(v_{ji}, a_{hi})) \qquad (3)$$

Figure 2(a) depicts the details of **overlap compensating neuron** (OCN), which represents a hyperbox of size equal to the overlapping region between two hyperboxes. OCN is active only when the test pattern falls in the overlapping region.

The activation function is given by equation (4) and (5).

$$d_{j_1} = U\left((\frac{1}{n}\sum_{i=1}^{n} 1 - f(a_{hi}-w_{ji}) - f(v_{ji}-a_{hi})) - 1\right) \times \left(-1 + \frac{1}{n}\sum_{i=1}^{n} max\left(\frac{a_{hi}}{w_{1ji}}, \frac{v_{1ji}}{a_{hi}}\right)\right) \qquad (4)$$

$$d_{j_2} = U\left((\frac{1}{n}\sum_{i=1}^{n} 1 - f(a_{hi}-w_{ji}) - f(v_{ji}-a_{hi})) - 1\right) \times \left(-1 + \frac{1}{n}\sum_{i=1}^{n} max\left(\frac{a_{hi}}{w_{2ji}}, \frac{v_{2ji}}{a_{hi}}\right)\right) \qquad (5)$$

where d_{j1} and d_{j2} are outputs for Class1 and Class2. V,W: min-max point of OCN. V_1,W_1,V_2,W_2: min-max point of overlapping hyperboxes $U(x)$: is a unit step function. The threshold used in the unit step function actives neuron only if the test sample is within the overlapping region. $f(x)$ is same as eq.(2) with $\gamma=1$.

The activation function of this neuron is such that it protects the class of the min-max point of the overlapping hyperboxes, which improves the learning accuracy. The membership grade decreases from point C to B gradually for class 1 and from B to C for class2 (Refer Figure 2b). The output of this neuron is connected to the two class nodes of overlapping classes (OC section Figure 1).

The containment compensation neuron (CCN) is shown in figure 2(c). This represents a hyperbox of size equal to the overlap region between two classes as shown in figure 2(d). This neuron is also active only when the test sample falls inside the overlapped region. The activation function of CCN is:

$$Oc_j = -1 \times U\left(\left(\frac{1}{n}\sum_{i=1}^{n} 1 - f(a_{hi} - w_{ji}) - f(v_{ji} - a_{hi})\right) - 1\right) \tag{6}$$

where Oc_j: output , V,W: min-max point of CCN, $U(x)$: unit step function, $f(x)$: same as eq.(2) with Υ=1

This function allows a hyperbox of one class to be contained in a hyperbox of different class. The output of CCN is connected to the class that contains the hyperbox of other class (CC section Figure1).

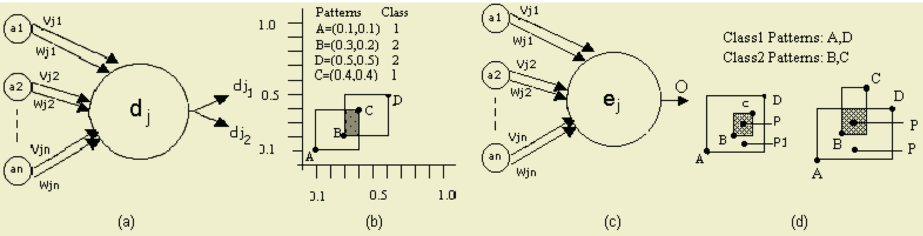


Fig. 2. Overlap and Containment Compensating Neuron

The number of output layer nodes in CL section is same as the number of classes learned. The number of class nodes in the CC, OC section depends on the nature of overlap the network faces during the training process. The final membership calculation is given by,

$$\mu_i = \max_{j=0}^{m}(b_j u_{ji}) + \min\left(\min_{j=1,..k}(d_j y_{ji}),\ \min_{j=1,..k}(e_j z_{ji})\right) \tag{7}$$

where U,Y,Z are the connection matrices for the neurons in the three sections.

Equation (7) takes care of multiple class overlaps. It gives maximum grade to a class from the available grades considering its compensation.

3 Learning Algorithm and Recall Procedure

The GFMCN learning algorithm creates and expands hyperboxes depending on the need of the problem. It tries to label the unlabeled data using the currently learned structure from the applied data. If there is any overlap, containment created (between hyperboxes of different classes) while expanding labeled hyperboxes, respective compensatory neuron is added to the network. Contraction procedure [5] is executed if there is any overlap between the unlabeled hyperboxes. Note that GFMN removes the overlap of unlabeled hyperboxes with all other hyperboxes in the network. Here we introduced the change and allow the unlabeled hyperbox to overlap with labeled hyperboxes. Moreover we assign its class label to the unlabeled hyperbox.

3.1 Learning Algorithm

Learning algorithm consists of mainly two steps namely Data Adoption and Overlap Test. **Assume** $\{a_h, C_i\}$ is a training data, $\{b_j, C_j\}$ a hyperbox for class C_j. Θ': current hyperbox size Θ_{max}: maximum hyperbox size. Initialize the network with b_1 having

$V_1 = W_1 = a_h$ and class C_i for an ordered pair of data $\{a_h, C_i\}$, Repeat the following steps 1 and 2 for the all-training patterns.

STEP 1: Data Adoption

Find a $\{b_{j,} C_j\}$ such that $(C_j = C_i$ or $C_j = C_0$ offering largest membership, $\Theta_J' \leq \Theta_{max}$ and is not associated with any OCN or CCN. Adjust the min-max points of hyperbox *bj* as,

$V_{ji}^{new} = \min (V_{ji}^{old}, a_{hi}) W_{ji}^{new} = \max (W_{ji}^{old}, a_{hi})$ where $i = 1,2 \ldots n$ (8)

and If $C_j = C_0$ *and* $C_i \neq C_0$ *then* $C_j = C_i$. Take a new training data point.

If no b_j is found create a new hyperbox with $V_j = W_j = a_h$ and class C_i.

STEP 2: Overlap Test

Assuming that hyperbox $b_{j,}$ expanded in previous step is compared with b_k.

If $C_j, C_k = C_0$ then do the overlap and contraction test as explained in [5].

Otherwise do the following tests:

a) Isolation Test: **If** $(V_{ki} < W_{ki} < V_{ji} < W_{ji})$ *or* $(V_{ji} < W_{ji} < V_{ki} < W_{ki})$ is true for any *i*, $(i \in 1 \ldots n)$ then $(b_{k,}, b_j)$ are isolated and take new input for training.

Else go for Containment test.

b) Containment Test: **If** $(V_{ki} < V_{ji} < W_{ji} < W_{ki})$ *or* $(V_{ji} < V_{ki} < W_{ki} < W_{ji})$ is true for any *i*, $(i \in 1,2 .. n)$ then

case1: if $C_j = C_0$ then assign $C_j = C_k$ go to step 1.

case2: if $C_k = C_0$ then assign $C_k = C_j$ go to step 1

case3: for $C_j, C_k \neq C_0$ create a CCN with hyperbox min-max co-ordinates given by

$V_{ci} = \max (V_{ki}, V_{li})$, $W_{ci} = \min (W_{ki}, W_{li})$ for i=1,2…n (9)

Else hyperboxes are not facing containment problem go to step (c)

c*) Overlap Compensation Neuron Creation:*

case1: if $C_j = C_0$ then assign $C_j = C_k$ go to step 1.

case2: if $C_k = C_0$ then assign $C_k = C_j$ go to step 1

case3: for $C_j, C_k \neq C_0$ create a OCN with hyperbox min-max co-ordinates given by

$V_{oci} = \max (V_{ki}, V_{li})$, $W_{oci} = \min (W_{ki}, W_{li})$ for i=1,2…n (10)

Avoid further expansion of hyperboxes belonging to different classes, which are facing the problem of overlap and containment, in the next expansion cycles.

3.2 Recall Procedure

Recall procedure is very simple. Class nodes in the each section calculate the class memberships and respective compensations. The summing node in the classifying neuron does the final grade calculation. The membership grade is computed according to equation 7.

4 Results

The GFMCN and GFMN algorithms can operate in three different modes i.e. pure classifier, clustering and hybrid. The performance of GFMCN was compared with that of GFMN in all the three modes. The datasets such as Iris, Wine and Ionosphere were used for the experiments.

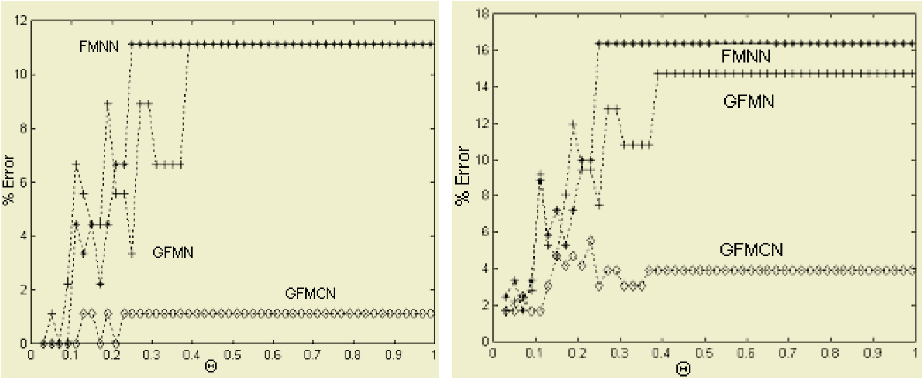
A) Performance in Classifying Mode

1. Real Data Sets: GFMCN algorithm was tested on widely used data sets such as Iris, Wine and Ionosphere. Table1 shows the results for the GFMCN compared to the results of GFMN, FMNN reported in [5], where all the patterns were used for training and testing and expansion coefficient was varied from 0-1 in step of 0.01.

2) Iris Data with Variation in Expansion Coefficient: Here the GFMCN, GFMN, FMNN algorithms are tested on fixed training data size (i.e. 60% of total patterns available in Iris data set). Data points were selected randomly for training. Expansion coefficient was varied in steps of 0.02, from 1.0 to 0.0. Here also, the learning error of GFMCN is almost independent of the expansion coefficient and the overall performance on Iris data set is better than GFMN, FMNN, see Figure 3 (a),(b).

Table 1. Percentage Recognition on Real Data

Data	Iris		Wine		Ionosphere	
	Max	Min	Max	Min	Max	Min
FMNN	97.33	92	100	94.32	98.01	84.77
GFMN	100	92	100	88.64	98.68	90.07
GFMCN	100	98	100	100	100	96.02



(a) Error for Training Data set

(b) Overall Recognition for Iris data Set

Fig. 3. Percentage error w.r.t. Θ

B) Performance in Hybrid Mode

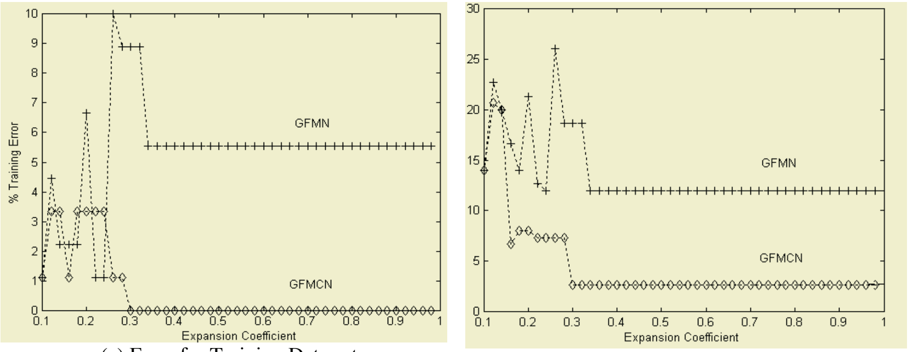
Training data set is generated by randomly selecting 60% Iris data from each class and 50% of the selected data was unlabeled for each class. Complete labeled Iris data set was used for testing so that we can evaluate the performance in hybrid mode. From Figure 4 (a) and (b) it is clear the GFMCN performance is better in training and testing even though Θ was varied from 1 to 0.1. Due to the inclusion of CNs and the new learning approach the unlabeled data is handled more efficiently and is labeled more correctly.

C) Clustering Mode

For the Clustering mode the performance of GFMN and GFMCN is same. There won't be any creation of CNs in this mode and hyperboxes will under go the same contraction procedure followed by GFMN [5].]

5 Conclusion

The proposed GFMCN learns the data on-line, in single pass. The compensatory neurons can handle the overlaps and containments of the hyperboxes more efficiently. From the experiments, we conclude that GFMCN can avoid the dependency of the network performance on Θ to a large extent hence single pass through learning is possible. As performance of GFMCN on labeled data is better than GFMN, its performance in hybrid mode (where mixture of labeled and unlabeled data used for training) is also better.



(a) Error for Training Data set

(b) Recognition for Iris data Set

Fig. 4. Percentage error for Hybrid mode w.r.t. Θ

References

1. L.A.Zadeh : Fuzzy Sets , Information and control, (1965), Vol. 8, 338-353.
2. P.K.Simpson : Fuzzy Min-Max Neural Network – Part I: classification ,IEEE Tran. Neural Networks, Sep. (1992), vol.3, no.5, 776-786.
3. P.K.Simpson : Fuzzy Min-Max Neural Network – Part II: Clustering , IEEE Tran. Fuzzy System, Feb. (1993), Vol.1, no.1, 32-45.
4. G.A.Baitsell : Human Biology , second edition, Mc-Graw Hill Book co. inc. NY, 1950.
5. B.Gabrys and A. Bargiela : General Fuzzy Min-Max Neural Network for clustering and Classification , IEEE Tran. Neural Network, May (2000). Vol.11, 769-783.
6. C.Xi, J.Dongming and L. Zhijian : Recursive Training for Multi-resolution Fuzzy Min-Max Neural Network Classifier , 6th Int. Cnf. Solid-State and Integrated Circuit Technology proceedings,(Shanghai), Oct. (2001), 131-134.
7. A.Rizzi, M. Panella, F.M. FrattaleMascioli: Adaptive Resolution Min-Max Classifiers, IEEE Trans. on Neural Networks, March (2002), Vol.13,pp. 402-414.
8. S.Abe and M.S. Lan : A Method for Fuzzy Rules Extraction Directly from Numerical Data and Its Application to Pattern classification , IEEE Trans. on Fuzzy Systems, Feb. (1995) Vol 3, No.1, 18-28.

9. G.Carpenter and S. Grossberg : A Massively Parallel Architecture for a Self-organizing Neural Pattern Recognition Machine , Computer Vision, Graphics & Image Understanding, (1987),Vol. 37, 54-115.
10. G.Carpenter, S. Grossberg, D.B. Rosen: Fuzzy ART: An Adaptive Resonance Algorithm for Rapid, Stable Classification of Analog Patterns, Int. joint Cnf. Neural Networks, IJCNN-91(Seattle), (1991), Vol.2, 411–416.
11. A.V. Nandedkar and P.K. Biswas: A Fuzzy Min-Max Neural Network Classifier with Compensatory Neuron Architecture, 17th Int. Cnf. on Pattern Recognition (ICPR2004) Cambridge UK, Aug., (2004), Vol. 4**,** 553-556.

An Analysis on Accuracy of Cancelable Biometrics Based on BioHashing

King-Hong Cheung[1], Adams Kong[1,2], David Zhang[1],
Mohamed Kamel[2], Jane You[1] and Toby, and Ho-Wang Lam[1]

[1] Department of Computing, The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
{cskhc,cswkkong,csdzhang,csyjia,cshwlam}@comp.polyu.edu.hk
[2] Pattern Analysis and Machine Intelligence Lab, University of Waterloo,
200 University Avenue West, Ontario, Canada
mkamel@uwaterloo.ca

Abstract. Cancelable biometrics has been proposed for canceling and re-issuing biometric templates and for protecting privacy in biometrics systems. Recently, new cancelable biometric approaches are proposed based on BioHashing, which are random transformed feature-based cancelable biometrics. In this paper, we consider the accuracy of one of the cancelable biometrics based on BioHashing and face. Through this analysis, as an illustration, we would like to raise an issue to be considered in cancelable biometrics: accuracy may be traded for biometrics being cancelable.

1 Introduction

Biometric authentications are increasingly performed under unattended and/or over networked environments. Attackers can attack via the exposed communication channels. [2], [3]

Ratha et al. [6] introduce the concept of cancelable biometrics to protect privacy in biometric authentication systems. It is achieved through intentional and repeatable distortions (or transformations) on biometrics in either the signal domain or the feature domain. The distortions for cancelable biometrics are ideally noninvertible. Nevertheless, the distortions can be invertible in practical use.

Ratha et al. [6] have given some example transforms for cancelable biometrics. As invertible examples, grid morphing and block permutation are offered and as a non-invertible example, high order polynomial is offered.

Cancelable biometric templates are essential for biometric authentication systems, especially for those operated under unattended and/or over networked environments. [6], [8]

Recently, variants [1], [5], [10] of BioHashing [11] which consists of feature domain random transformation and discretization, were extended as a means to cancelable biometrics.

In this paper, we bring out the issue that accuracy is traded for biometrics being cancelable through a case study on BioHashing and face. This issue is important to cancelable biometrics and thus biometric authentications. The rest of this paper is organized as follows. Section 2 presents the issue that should be considered. Section 3 presents an empirical test to demonstrate the issue. Section 4 offers our conclusions.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1168–1172, 2005.

2 Cancelable Biometrics

We have briefly reviewed the major reasons and concepts of cancelable biometrics in the above section. In this section, we would like discuss the issue in cancelable biometrics including accuracy in section 2.1. Accuracy is the short term for recognition accuracy of a biometric system.

2.1 Accuracy

To make biometric system practical, satisfactorily high recognition accuracy is required. Therefore, we have to make sure the recognition is accurate enough for an application. Noninvertible transforms in cancelable biometrics for both signal and feature domain can lead to information loss that affects the discriminating ability and results in deterioration of accuracy. Invertible transforms can replace noninvertible transforms to avoid information loss.

Since cancelable biometrics is matched in the transformed domain, we have to define a suitable feature extraction as well as similarity measure in the transformed domain. Otherwise, the accuracy may not be guaranteed. For example, if a monotonically decreasing function is the transform and L_2-norm is the similarity measure, it will be inaccurate to measure the similarity in the transformed domain using L_2-norm. We, therefore, have to look for transforms that keep meaningful relationships for feature extraction (for signal domain transforms) and similarity measure (for both domain) afterwards. Certain transforms may still be used for providing cancelable biometrics even though we cannot find a suitable strategy to maintain the original recognition accuracy.

Connie et al. [1], Pang et al. [5] and, Teoh and Ngo [10] presented prototypes of cancelable biometrics for palmprint and face based on BioHashing [11]. BioHashing consists of two major steps, feature domain random transformation, and discretization (a two level quantization). Their transform is a kind of noninvertible feature domain transform. As the transform is noninvertible, the raw template can be better protected. Random transform is one of the viable approaches to provide cancelable biometrics. Nonetheless, we believe that noninvertible random transformations will destroy the optimality of most feature representations and thus the recognition accuracy deteriorates. There is a tradeoff in the feature domain between optimality in representation and similarity matching and biometrics being cancelable, i.e. the error rate of authentication increases [12].

3 Analysis on Accuracy of a BioHashing Based Cancelable Biometric

In order to ensure cancelable biometrics is practical, we have to look at the system performance. We understand there is tradeoff between cancelability and recognition accuracy. So, we now look at the amount of recognition accuracy that can be traded for biometrics being cancelable by a test of a cancelable biometric for face based on BioHashing proposed by Teoh and Ngo in [10].

Wavelet Fourier Mellin Transform (WFMT) [4], [9] is the feature extraction technique used in [10]. The ORL face database [7] is a well-known public face database and is adopted in [10] and Lai et al. [4]. There are 10 different images for each of 40 distinct subjects. For some of the subjects, the images were taken at different times, varying lighting slightly, facial expressions (open/closed eyes, smiling/non-smiling) and facial details (glasses/no-glasses). All the images are taken against a dark homogeneous background and the subjects are in up-right, frontal position (with tolerance for some side movement). The size of each image is 92×112 of 8-bit grey levels.

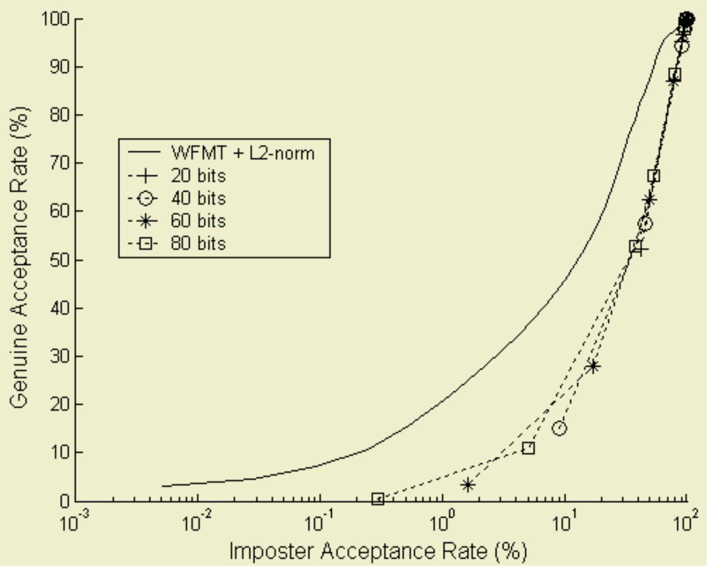


Fig. 1. Test on accuracy of various BioCodes of BioHashing compared to Euclidean distance

The implementation details of WFMT are listed in Table 1 and thresholds (for quantization) used for various bits of BioHashing are listed in Table 2 for reference. The thresholds chosen are the same as indicated in [10], [11]. Moreover, the number of matching for estimating the genuine and impostor distributions based on the Teoh and Ngo's matching scheme, which matches only the n^{th} image of one person to the n^{th} image of other persons to determine the impostor distribution where $n = 1\ldots10$, are 1800 (i.e. [(1+9)×(9/2)]×40) and 7800 (i.e. [(1+39)×(39/2)]×10) respectively. (for details please refer to [10])

The Receiver Operating Characteristic (ROC) curves [2], [3] of WFMT with BioHashing of 20, 40, 60 and 80 bits are plotted (dashed lines) along with WFMT with Euclidean distance, L_2-norm (solid line) in Figure 1. Nearest-Neighbour-Classifier is used to determine the matched identity for WFMT with BioHashing and Euclidean distance. From Figure 1, the optimality of feature representation is shown to be destroyed by the noninvertible random transform and quantization and thus the recognition accuracy deteriorates. The performance of WFMT with BioHashing is even worse than that of WFMT with Euclidean distance, i.e. dashed lines are beneath the solid line.

Table 1. The details of WFMT implementation

Processes/Variables/Parameters	Values/Descriptions
Raw image sizes	92×112, no preprocessing
Wavelet	db7
Level of Wavelet Decomposition	1
Wavelet transformed image sizes (LL band)	52×62
Log-polar transformation	Largest inscribed circle, bicubic interpolation, 62 logarithmic levels
Highpass Filter	same as in [10], $H(x,y) = (1-\cos(\pi x)\cos(\pi y))\times(2-\cos(\pi x)\cos(\pi y))$

Table 2. Thresholds used for various bits of BioHashing

Bits	Thresholds
20	0
40	0
60	0
80	0

4 Conclusions

We have presented a brief review of cancelable biometrics. We have raised an issue in cancelable biometrics worth for consideration. Through an analysis of accuracy of an existing approach to cancelable biometric, it is shown that biometrics being cancelable is not free lunch. The accuracy can be traded because we are not able to find a strategy to integrate the transform, feature extraction and similarity measure as a whole.

Acknowledgement

The authors would like to thank for the partial support of the research grants from Hong Kong Government (UGC) and the research grants for iJADE projects PG50 and A-PF74 of The Hong Kong Polytechnic University.

References

1. Connie, T., Teoh, A., Goh, M., Ngo, D: PalmHashing: a novel approach to cancelable biometrics. Information Processing Letter **93** (2005) 1-5
2. Jain, A., Bolle, R., Pankanti, S. (eds.): Biometrics: Personal Identification in Networked Society. Kluwer Academic Publishers, Boston Mass (1999)
3. Jain, A.K., Ross, A., Prabhakar, S.: An Introduction to Biometric Recognition. IEEE Transactions on Circuits and Systems for Video Technology **14** (2004) 4-20
4. Lai, J.H., Yuen, P.C., Feng, D.C.: Face recognition using holistic Fourier invariant features. Pattern Recognition **34** (2001) 95-109
5. Pang, Y.H., Teoh, A.B.J., Ngo, D.C.L.: Palmprint based cancelable biometric authentication system. International Journal of Signal Processing **1** (2005) 98-104
6. Ratha, N.K., Connell, J.H., Bolle, R.M.: Enhancing security and privacy in biometrics-based authentication systems. IBM Systems Journal **40** (2001) 614-634

7. Samaria, F., Harter, A.: Parameterisation of a stochastic model for human face identification. Proceedings of the 2nd IEEE Workshop on Applications of Computer Vision, Sarasota (Florida) (1994) 138-142 [paper and ORL face database both available online at http://www.uk.research.att.com/facedatabase.html]
8. Schneier, B.: The Uses and Abuses of Biometrics. Communications of the ACM **42** (1999) 136
9. Srinivasa Reddy, B., Chatterji, B.N.: An FFT-Based Technique for Translation, Rotation, and Scale-Invariant Image Registration. IEEE Transactions on Image Processing **5** (1996) 1266-1271
10. Teoh, A.B.J., Ngo, D.C.L.: Cancellable biometerics featuring with tokenised random number. To appear in Pattern Recognition Letters (2004) [available online but obtained from A.B.J. Teoh]
11. Teoh, A.B.J., Ngo, D.C.L, Goh, A.: BioHashing: two factor authentication featuring fingerprint data and tokenised random number. Pattern Recognition **37** (2004) 2245-2255
12. Uludag, U., Pankanti, S., Prabhakar, S., Jain, A.K.: Biometirc Cryptosystems: Issues and Challenges. Proceedings of the IEEE **92** (2004) 948-960

Condition Monitoring Capability Developed Through a Knowledge Transfer Partnership Between a Small Company and a University

Robert Howlett[1], Gary Dawe[1], and Terry Nowell[2]

[1] Knowledge Transfer Partnership Centre and Intelligent Systems and Signal Processing
Research Group, University of Brighton, Moulsecoomb, Brighton, BN2 4GJ, UK
R.J.Howlett@brighton.ac.uk
[2] Powerlase Ltd, Imperial House, Link 10, Napier Way, Crawley, West Sussex, RH10 9RA, UK

Abstract. This paper describes the benefits that small companies can obtain from the United Kingdom Department of Trade and Industry Knowledge Transfer Partnerships business support product. This provides funding for collaborations between universities and companies in order that knowledge and technology transfer between the partners can occur. Knowledge Transfer Partnerships can support a wide range of commercial activities, proving there is a solid business case to show that the government support will lead to improvements in the profitability of the company. This paper describes the Knowledge Transfer Partnerships product. It then goes on to present a case study of a business collaboration under the scheme between Powerlase Ltd, a manufacturer of laser systems, and the University of Brighton, to implement intelligent condition monitoring in the Company's products.

1 Introduction

In 2003 the Lambert Report [1] identified that in the UK, public spending on the teaching of university students was over £3billion per annum and the amount spent on research in universities was over £2 billion per annum. Increasing attention is being given to a third source of income, arising from commercial activities and income generation. This additional ("third-stream") mode of financing is of growing importance. Lambert defined "Transferring the knowledge and skills between universities and business and the wider community" as *Knowledge Transfer* although the term had been in use for some time previously. He noted that knowledge transfer provides a number of benefits, including that it increases the economic and social return on public investment.

The Knowledge Transfer Partnerships (KTP) programme [2] is a business support product operated by the UK Department of Trade and Industry (DTI). KTP has gained an enviable reputation for delivering high-quality innovation to UK companies through three-way knowledge-transfer interactions between firms, universities and skilled graduates.

KTP and its predecessor (the Teaching Company Scheme, TCS) have together been in operation since the 1975. There are currently (in 2005) about 850 Partnerships in operation. There is the intention to grow to 1200 Partnerships, and funding is available to support this number.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1173–1179, 2005.

Each Knowledge Transfer Partnership involves three components, a UK company, a Knowledge-Base Partner, and a graduate, called a KTP Associate. The company must have a need for a demanding project of a strategic nature. This must be something that will lead to real business benefits in terms of increased turnover and profit, or safeguarded market-share. The project must also be something that the company could not do for itself, without the help of the Knowledge-Base Partner.

The Knowledge-Base Partner is most commonly a university. The Knowledge-Base Partner must have a high level of skills and expertise to contribute to the project. This is provided through an Academic Supervisor who has technical skills in the area of the project, and who also mentors the KTP Associate. To be suitable for KTP, the Associate must be able to benefit from the associate development programme that is offered, which means they must usually have qualified in the recent past few years. They must have a qualification appropriate to the project, for example, a first or upper second class honours degree for a university-company KTP.

During the operation of the KTP, the Associate works in the company carrying out the project. The Associate works under the direction of the Academic Supervisor, and a member of the company staff, the Industrial Supervisor, who acts as the Associate's line manager. The Academic Supervisor visits the company on a regular basis, and commits to contributing half a day a week of their time to the project over the life of the partnership. The Associate works under the company's conditions of service, although they have a contract with the university partner, who is given responsibility for managing the grant. About 70% of Associates join the company at the end of the Partnership, although there is no compulsion on the company to offer a permanent position, and no compulsion on the Associate to stay.

Each Knowledge Transfer Partnership carries attractive funding from the DTI or one of a number of other sponsors to the scheme. The total project budget is currently about £80k over two years. If the company is a Small to Medium Enterprise (SME), approximately within the European Union definition, i.e. has fewer than 250 employees and turnover and company values within certain limits, a Knowledge Transfer Partnership provides funding of 60% of the project budget, and the company pays the remaining 40%. If the company does not qualify as an SME the project attracts about 40% funding, and the company contribution is 60%.

2 Benefits to the Company

In order to obtain a Knowledge Tranfer Partnership and the grant income it includes, a credible and financially beneficial business case must be presented in the proposal document. At the end of the project the benefits arising from the project are assessed by independent reviewers. The business benefits actually obtained vary widely because of the wide range of types of projects, companies and business sectors. National statistics [1] indicate that typically, the following gains due to the Knowledge Transfer Partnership are achieved: an increase in the overall value of the business by 52%; an increase in sales or turnover of 46%; and an increase in profitability of 42%.

In concrete terms, on average, KTP projects achieve an increase in annual profits of over £170k and the creation of four new jobs.

The Knowledge Transfer Partnership can be focussed on any one of a number of business objectives. For example, new product development, manufacturing process improvement, revised business strategy or challenging new markets, are all valid objectives for a KTP.

3 The University of Brighton KTP Centre

The University of Brighton KTP Centre was formed in 1993 as Brighton TCS Centre, the first single-university centre formed to promote, set up and operate TCS Programmes (TCS then being the equivalent to KTP). In the 12 years it has been in existence the KTP Centre has attracted approximately £6.8 million in funding and has operated programmes with approximately 80 KTP Associate. The companies the KTP Centre works with come from a number of industry sectors, which mainly reflect the areas of the University that have expertise to offer in the domains that attract KTP funding. The departments in the University that have provided the greatest support to KTP are the Schools of Engineering, Pharmacy and Bio-molecular Sciences, Environment, Computing Mathematics and Information Science and the Centre for Research in Innovation Management

4 Case Study: Condition Monitoring Through Intelligent Systems

Powerlase Limited is a growing company, rising in prominence, that is located in the South East of England, but which trades internationally. The Company manufactures Diode Pumped, Solid State lasers developed for industrial applications in the materials processing and microelectronics market places. Powerlase has combined high average power with high intensity nanosecond pulses, not previously available together, to achieve high process throughput in a range of materials processing. This cutting edge technology currently serves applications in Aerospace, Automotive, Microelectronics, PCB production, Ablative Lithography, and many others. To further enhance the already good reliability of its products Powerlase wish to implement proactive condition monitoring to achieve prediction of failures and avoid unplanned downtime. To do this the Company wished to take advantage of the expertise of the Intelligent Systems and Signal Processing (ISSP) Research Group at the University of Brighton. The Research Group had a proven track record of applying neural, fuzzy and expert systems techniques to the solution of industrial condition monitoring problems. The KTP scheme provided funding for a KTP Associate to work with Powerlase to develop intelligent condition monitoring tools and techniques to be applied to its products.

4.1 The Rationale for Condition Monitoring

The motivation for the implementation of a condition monitoring system is provided by a compelling business case, that can be summarised as follows:-

- The company can win a truly competitive edge and meet its ambitious growth targets if it can show that the demonstrable performance advantages of its lasers are further enhanced by high levels of reliability. The benefits to the company are higher customer satisfaction and less field support expenditure.

- Today's customers have high expectations of new equipment. Good reliability measured in high Mean Time Before Failure (MTBF) and the ability to schedule downtime of equipment are all desirable in the purchasing decision of new plant.
- Products entering a new marketplace, as is the case in Powerlase, need to be able to demonstrate results that clearly show the long-term reliability of its systems. Part of the route to high MTBF is condition monitoring of the equipment. By gathering data via sensors distributed about the laser early identification of problems can be made, this also provides the ability to predict faults before they occur and enable planned downtime. Using data gathered, advanced process control is enabled to reduce product wastage and rejection rates.
- Any unplanned machine/tool breakdowns, or maintenance, impacts on profitability via missed production deadlines and product quality. The laser is usually integrated into a machine tool for material processing. If the laser stops then the whole production line stops.

To summarise, the benefits of condition monitoring are improved reliability, reduced repair time and costs, avoided revenue loss, and improved safety as well as the ability to accurately forecast production capabilities.

4.2 Condition Monitoring of Laser Systems

There is relatively little published in the literature on research into condition monitoring of laser systems. In particular, intelligent systems have not been widely used. The methods employed are mathematical or model based.

A methodology has been described [3] to discriminate electro-optical failure signatures related to an optical alignment drift in the laser module in comparison of those related to the gradual change of electro-optical parameters of the laser diode. The application was low powered devices used in communications systems. The methodology allowed the extraction of three main failure signatures after 300 thermal cycles. Cordera et al. [4] outline two strategies to monitor the laser output at the final turning mirror of a high power CO_2 laser. The development of a dynamic model for the condition monitoring of an industrial CO_2 laser has been explained [5]. There is little description of strategic condition monitoring in the literature. However, a strategy to condition monitor the cooling system of a CO_2 laser and then estimate useful laser lifetime, has been reported [6]. Identification of the fault-prone components is conducted followed by selection of indicator variables and mapping these variables to faults. Statistical methods are then used to predict the useful lifetime of the laser system based on the measured variables and system inputs. The strategy is then tested on a system and verified using simulation.

4.3 A Strategy for Intelligent Condition Monitoring

The laser system that formed the basis of this was a Powerlase Starlase AO4 high power, solid state, pulsed Neodymium Yttrium Aluminium Garnet (ND:YAG) laser. The system has an average power output of 400W at a 10 kHz pulsed repetition rate with an M^2 (indictive of beam quality) of 22. The system consists of a control rack that houses the laser controller, switching signal generators, diode power supplies and water-cooling heat exchanger. The Laser enclosure, which can be mounted up to 7 meters away from the rack, contains the optical laser cavity, sensors for control feedback and safety interlocks. Most of the optical components are liquid cooled using a sealed de-ionised water system.

The condition monitoring philosophy [7] that was developed during the project was to:-

- identify the system variables that were indicative of failure;
- devise appropriate monitoring sub-systems;
- apply intelligent neural data analysis algorithms;
- utilise rule-based inference mechanisms to diagnose the fault condition.

The project followed a structured approach to identify variables that need to be monitored. A systems level model of the laser system was created using the Unified Modeling Language (UML) to aid in understanding the system. The model mapped objects to the components and sub-systems and identified methods and attributes of the components. The inter-connections were then added to create a complete but simplified graphical model of the system. Usage scenarios and task flow were then modeled to identify the system functionality.

A diagnostic matrix was created to map possible and observed failures in the laser system to influences causing these effects. The matrix was then scored according to the likelihood or impact of the failure on system performance. Priority placement of measurement sensors on the higher risk components/areas could then be implemented. As well as redesign of key components.

Custom logging software to store the variables into files was developed in C. This enabled data to be gathered at customer sites, as well as from in-house system testing, for later analysis. Analysis of the data provided the evidence that identified a possible use of a neural network for on line analysis of the laser pump module.

4.4 Lasing Module Tests Using a Neural Network

A phase of the condition monitoring system development involved a feasibility study of a neural-network based testing technique. This was to determine whether the characteristics of the lasing modules were within specification prior to their integration into complete systems and without the requirement for a full laser.

The lasing module was mounted into a short-cavity test rig. The power-current characteristic of the lasing module was recorded. This was done by making 14 measurements of current and power in 2 Amp steps from 26 Amps to 52 Amps. Data was obtained in the form of characteristics when the lasing modules were new, and when they were returned from the field as faulty. This resulted in 23 sets of data, 15 corresponding to lasing modules judged to be in-specification and eight corresponding to modules where power output problems had been reported out-of-specification. Figure 1 shows typical characteristics for the two categories of lasing modules.

For this feasibility study a C language implementation of the Multi-layer Perceptron (MLP) neural network using the back-propagation training algorithm was chosen as the MLP is known to be an efficient pattern discrimination tool [8]. It is acknowledged that theoretically the number of training examples required for good generalisation is approximately 10 times the number of weights. However, experience has

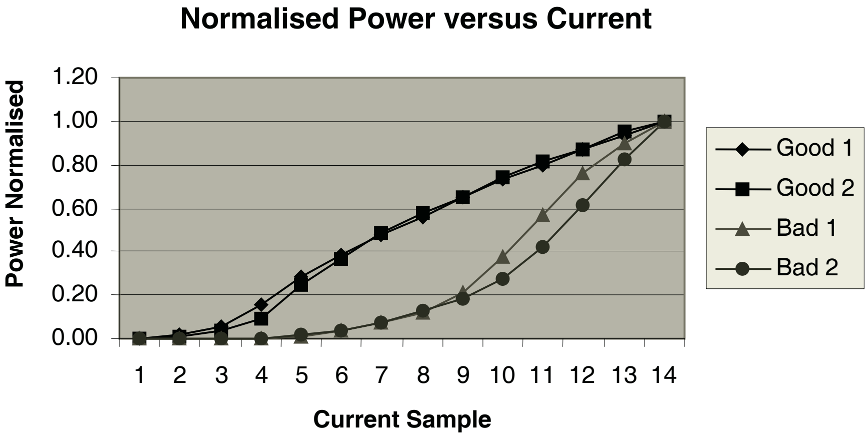


Fig. 1. Power vs current characteristic for in-specification (good) and out-of-specification (bad) lasing modules

shown that the likelihood of obtaining good results with large data sets can be tested with smaller amounts of data [9]. The neural network parameters used were as follows:- 14 input nodes, five hidden nodes and two output nodes.

Once trained the network was then presented with 10 sets of unseen data. The test data consisted of 10 sets of normalised data divided into five in-specification and five out-of-specification sets. The network performance was then noted. It was found that in recall the neural network achieved a 100% correct classification rate. This demonstrated the feasibility of using the neural network testing technique to quickly and conveniently differentiate between in-specification and out-of-specification lasing modules. A complete account of the experimental work and results is available [10].

5 Conclusion

UK universities find it increasingly important to access third-mission funding to supplement incomes from teaching and research. In addition there are a range of additional benefits arising from interaction with industry. Examples include increasing the relevance of teaching, the opportunity to put theoretical research skills into practice and experience of operating in a commercial environment. Companies can gain commercial benefits from forming partnerships with universities. The advantages include tapping in to research skills and expertise, both at a specific technical level, and also on a more peripheral level, for example, an engineering project that also makes use of university knowledge of business development. Graduates can benefit from forming the facilitator of knowledge transfer between universities and companies. They gain experience, develop skills in project management, make technical progress, and become able to take ownership of an important strategic project. In the UK, KTP is a prominent knowledge transfer mechanism. KTP allows companies, universities and graduates to benefit, forming a truely win-win-win situation.

References

1. Lambert, R. The Lambert Review of Business-University Collaboration Interactions. Her Majesty's Stationary Office, Norwich, UK. (2003) p. 31.
2. KTP Online Web Site. http://www.ktp-online.org.uk, (2005).
3. Deshayes, Y., Bechou, L., Mendizabal, L., Danto, Y. Early failure signatures of 1310 nm laser modules using electrical, optical and spectral measurements. Measurement 34 (2003) pp. 157–178.
4. Cordera, D.A., Evans, D.R., Tyrer, J.R., Freeland, C.M. & Myler, J.K. High Power Laser Beam Delivery Monitoring for Laser Safety. Optics and Lasers in Engineering 27 (1997) pp. 479-492
5. Tu, J.F., Katter, J.G., Monacelli, J.G., & Gartner, M. A dynamic model for condition monitoring of a high power CO_2 industrial laser. Journal of Dynamic Systems Measurement & Control 121 (1999) pp. 157-164
6. Koomsap, P., Prabhu, V.V., Shaikh, N.I. Schriempf, J.T. McDermott, J.H. Condition monitoring and lifetime estimation of a CO_2 laser Journal of Laser Applications Vol.15 No.4 (2003) pp. 285-293
7. Wang. K. Intelligent Condition Monitoring and Diagnosis Systems. IOS Press ISBN 1 58603 312 3. pp. 103-118
8. Haykin. S. Neural Networks A Comprehensive Foundation. 2nd Edition. Prentice Hall ISBN 0 13 908385 5. pp. 139.
9. Rumelhart, D.E. and McClelland, J.L. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol.1 (1986) MIT Press. Cambridge, MA. pp.148.
10. Dawe, G.P., Howlett, R.J., Joyce, I. and Nowell, T. Condition monitoring of high-power, Q-switched, diode-pumped lasers. British Institute of Non-Destructive Tesing Conference. Cambridge. (2005). Accepted for publication.

Extraction of Lexico-Syntactic Information and Acquisition of Causality Schemas for Text Annotation

Laurent Alamarguy[1], Rose Dieng-Kuntz[1], and Catherine Faron-Zucker[2]

[1] ACACIA, INRIA Sophia Antipolis
{Laurent.Alamarguy,Rose.Dieng}@sophia.inria.fr
[2] MAINLINE, I3S, Sophia Antipolis
faron@essi.fr

Abstract. We present the INSYSE method for the annotation of texts, based on extraction of semantic relations from syntactic structures. We apply this method to a corpus of 5000 Medline abstracts about central nervous system diseases and gene interactions. Our cooperative approach focuses on (1) extracting lexico-syntactic information from sentences in the corpus comprising causation lexemes and (2) elaborating unification grammar rules which enable to extract instantiated conceptual schemas from this information. They are translated into RDF annotations which used by the semantic search engine Corese to query the corpus about functions of genes and their correlations with particular diseases.

1 Introduction

The notion of causality is essential to understand some correlations in functional genomics. The automation of the detection of such causality correlations and their conceptual representation is a keystone to build a community memory. This can be achieved by using some Natural Language Processing (NLP) methods.

We propose a semi-automatic method of text annotation which is based on the acquisition of conceptual templates from the extraction of lexico-syntactic structures. We call it INSYSE (Interface of SYntax-SEmantics). It is applied to a corpus about 5000 biomedical abstracts from Medline, dealing with central nervous system pathologies and the gene interactions in these pathologies. We aim at generating semantic annotations on these abstracts to inform about gene functions and their causal relations with some diseases. A memory of the community of the actors in biomedical field can thus be built.

INSYSE only focuses on causation relation analysis, since the aim of detecting some correlation between gene functions and pathologies favors this focus and our corpus is characterized by numerous and various causation markers. However, some other relationships certainly underlie in the comprehension of these correlation, but we do not address their study in this work. We study *intra-clausal* causation markers; discourse markers, that may overlap several sentences, are out of the scope of our study, since their construal and processing require another linguistic analysis.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1180–1186, 2005.

In this paper, we introduce the various steps of the INSYSE method, as depicted in Figure 1. INSYSE stresses on the processing of a fine grained syntactic analysis (step 2), and the construal of an accurate syntax-semantic interface (step 3). The second stage mainly relies on the merging of a terminological extraction with a partial syntactic parsing, so as to provide domain-relevant concepts and accurate interconnections between these concepts. The syntax-semantic interface is based on a cognitive-functionalist approach [10] advocating a strong correlation between semantic roles and syntactic functions from prototypical mapping (active form) and from dynamic operations such as *perspectivization* enabling to construe passive or nominal form, or dative shift.

In section 2 we describe the extraction step of lexico-syntactic information through sentences containing some causation lexemes. Section 3 is dedicated to the elaboration of rules based on unification grammars which enable to extract some lexico-syntactic information peculiar to some instantiated conceptual schemas. In section 4, we describe how these schemas are translated into RDF(S)[1] annotations from which the corpus will be queried through the inference search engine CORESE [5], once a concept matching phase will have been processed. In conclusion, the INSYSE method is compared with other approaches related with text annotations and we sketch its on-going evaluation.

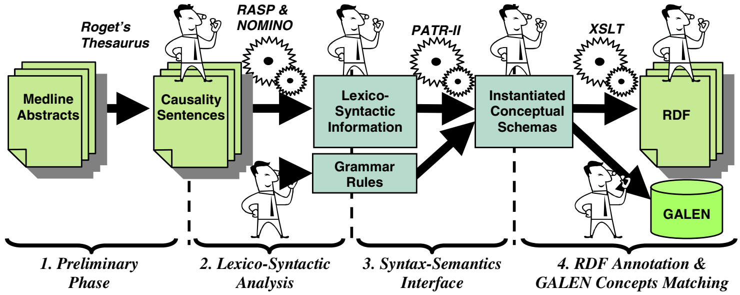

Roget's Thesaurus
RASP & NOMINO
PATR-II
XSLT
Medline Abstracts
Causality Sentences
Lexico-Syntactic Information
Instantiated Conceptual Schemas
RDF
Grammar Rules
GALEN
1. Preliminary Phase
2. Lexico-Syntactic Analysis
3. Syntax-Semantics Interface
4. RDF Annotation & GALEN Concepts Matching

Fig. 1. INSYSE in a Nutshell

2 Lexeme Extraction from Texts for a Lexicon Construction

The INSYSE preliminary step consists in the selection of relevant sentences, from abstracts in our corpus, so as to operate the lexico-syntactic analysis. It aims at identifying the sentences describing gene functions interacting in nervous system pathologies, and the relevant sentences are selected according to the causative lexemes they contain, such as *causing*, *triggering*, *activating*, etc. This stage is guided by the abstract relations of causation listed in the Roget's Thesaurus.

The syntactic analysis of the selected sentences is based on the application of the RASP shallow parser [3] on the whole corpus. So for each sentence, the

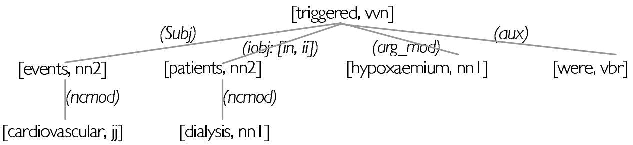
[1] http://www.w3.org/TR/1999/REC-rdf-syntax-19990222

syntactic functions of its lexemes are revealed. A dependency tree is built with a lexeme for each node and drawn from syntactic functions. RASP assigns to each lexeme the following lexico-syntactic information:

- syntactic dependency relations, *e.g.*, in NP *dialysis patients*, the noun *patients* symbolizes the 'head' and *dialysis* is the dependency of *patients*;
- grammatical relations such as *subject*, *object*, *auxiliary*, etc.;
- morphosyntactic tags (PoS tags) indicating the grammatical category of each word through context.

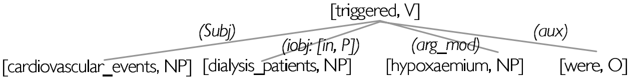
Let us consider the following excerpt of our training corpus:

Cardiovascular events were triggered in dialysis patients by hypoxaemia. The parsing of RASP construes the following dependency tree:



Events is construed as the sentence subject (*subj*), *triggered* as the verbal predicate (*head*), *patients* as the indirect object, specified by the preposition *in* (*iobj: in*), *hypoxaemia* as an adjunct (*arg_mod*), and *were* as the auxiliary (*aux*). *Cardiovascular* and *dialysis* are construed as modifier of respectively *events* and *patients. Hypoxaemia*, *dialysis*, *events* and *patients* are commoun nouns, singular (*nn1*) or plural (*nn2*), *triggered* is a past-participle verb (*vvn*), *were* is a preterit form of the *be* auxiliary (*vbr*), *cardiovascular* is a general adjective (*jj*) and *in* a preposition (*ii*).

The constructed lexicon is refined by comparing the lexical entry embodied by each lexeme with the automatic term extraction operated by Nomino [6] on the same corpus. This terminological extractor provides a more accurate syntagmatic categorisation enhancing the relevance of lexical entry to the domain vocabulary. So the revelation of coherent and relevant domain terms constitutes a fundamental step in semantic extraction from texts [2]. For instance, Nomino analysis extracts from the above sentence the term *dialysis_patient*; it will replace the RASP lexeme *patient* as a lexical entry. *Dialysis_patient* inherits the lexico-syntactic information of *patient*, which is the head of the nominal phrase (NP) *dialysis patients*: its dependency relation with *triggered* as an argument, its grammatical relation with *triggered* introduced by *in* (indirect object). The RASP dependency tree then becomes:



Thus, each lexical entry is constituted of lexicographical domain information, and morpho-syntactic information that will be processed by our grammar rules.

3 Instantiated Conceptual Schemas Acquisition

The second stage of INSYSE consists of acquiring conceptual schemas capturing the meaning of a sentence, from the lexico-syntactic information associated with each lexical entry extracted from parsed corpus sentences. To achieve it, we use the grammatical parser PATR-II [12] defined by a unification formalism, and enabling (1) to reveal a peculiar complex and coherent semantic structure from more primitive substructures, and (2) to construe *perspective* grammatical operations such as passivation or nominalization.

So, we have defined a set of about 50 grammar rules from the manual study of representative causation constructions in the training corpus. Based on feature unification and constraints, rules parse a sentence using the extracted lexico-syntactic information and build an instantiated conceptual schema. Thus, these rules embody the syntax-semantics interface, since they map syntactic functions such as *subject*, *object*, etc. with semantic functions like *agent*, *patient*, etc. The following five rules in Table 1, extracted from the grammar we have built and dedicated to the causation construal, parse passivation:

Table 1. Example of grammar rules processed by PATR-II

```
    Rule {Clause Passivation}
(1) S -> NP VP
(2)    <S sem pred> = <VP sem pred>
(3)    <VP sem postag> = VVN
(4)    <S AGT> = <VP sem arg2>
(5)    <S AGT sem case> = Arg_Mod
(6)    <S PAT> = <NP>
(7)    <NP sem case> = Subj
(8)    <S SET> =<VP sem arg1>


    Rule {Passive Predication}
(1) VP -> V2 PP1 PP2
(2)    <VP sem pred> = <V2>
(3)    <VP sem arg1> = <PP1 sem pred>
(4)    <VP sem arg2> = <PP2 sem pred>
```

```
    Rule {Passive Predication Operator }
(1) V2 -> O V
(2)     <V2 sem pred> = <V sem pred>
(3)     <V2 postag> =<V sem postag>
(4)     <V2 sem arg> = <O>
(5)     <O sem case> = Aux

    Rule {Periphery1}
(1) PP1 -> P NP
(2)     <PP1 sem pred> = <NP>
(3)     <PP1 sem arg> = <P>

    Rule {Periphery2}
(1) PP2 -> P NP
(2)     <PP2 sem pred> = <NP>
(3)     <PP2 sem arg> = <P>
```

The rule *Clause Passivation* refers to the passive form of a sentence S, constituted with a noun phrase NP and a verb phrase VP (1), and stipulating that:

- The semantic predicate of S will be inherited from VP (2);
- if VP is a past participle verb (3), NP is subject (7) and the *agent* role AGT is filled by the adjunct Arg_Mod (5), then NP fulfills the *patient* semantic role PAT of S (6), and the semantic argument arg2 of VP plays the AGT role of S (4);
- the semantic argument arg1 of VP plays the *setting* role SET of S (8).

The rule *Passive Predication* refers to the passive form of a VP predicative structure, constituted with a verbal structure V2, and two prepositional phrases PP1 and PP2 (1), stipulating that:

- the semantic predicate of VP will be inherited from V2 (2);
- the argument arg1 of VP will be inherited from the semantic predicate of PP1 (3);
- the argument arg2 of VP will be inherited from the semantic predicate of PP2 (4);

The rule *PassivePredication Operator* refers to the nucleus structure of a verbal constituent V2, constituted with an operator O and a verb V (1), and stipulating that:

- the semantic predicate of V2 will be inherited from the semantic predicate of V (2);
- the morpho-syntactic category of V2 will inherit the morpho-syntactic tag of the semantic structure of V (3);
- if the operator O is auxiliary (5), then O becomes argument of V2 (4).

The rules *Periphery1* and *Periphery2* refer to a PP prepositional structure, constituted with a preposition P and a noun phrase NP (1), both stipulating that:

- the semantic predicates of both PP1 and PP2 correspond to the nominal phrase NP (2), and the arguments of PP1 and PP2 correspond to the preposition P (3).

When processing the lexicon file extracted from the above sentence taken as example, these five rules parse the following conceptual schema through PATR-II:

```
[cat: S
   AGT:[cat: NP
            lex:Hypoxaemia                   sem:[case: Arg_Mod, pred: HYPOXAEMIA]
   PAT:[cat: NP
            lex: cardiovascular_events       sem: [case: Subj, pred: CARDIO-EVENT]]
   SET:[cat: PP
            lex: in_dialysis_patients        sem: [case: iObj, pred: DIALYSIS_PATIENT]]
   sem:[pred: [postag: VVN, pred: TRIGGER]]]
```

This schema stipulates that the *agent* of TRIGGER is fulfilled by HYPOXAEMIA, the *patient* of TRIGGER is fulfilled by CARDIO-EVENT and the *setting* of TRIGGER is fulfilled by DIALYSIS-PATIENT. Thus, semantic relations interconnect semantic predicates of lexemes extracted from the corpus.

Moreover, this conceptual schema would be also elaborated to construe an active or nominalized form, following the relevant rules.

4 Document Annotation from Conceptual Schemas

These acquired instantiated conceptual schemas will constitute semantic annotations of the Medline abstracts whose sentences have been parsed. This last stage aims at translating these schemas into the RDF semantic web standard language. The output of the PATR-II parsing in XML syntax is converted into RDF by using a XSLT style sheet. For instance, the above conceptual schema is translated into the following RDF annotation:

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:gal="http://www.sophia.inria.fr/acacia/galien#">
  <gal:Abstract rdf:about="http://www.sophia.inria.fr/acacia/medline#a324">
      <gal:hasForCausationSchema>
        <gal:CausationSchema rdf:about="http://www.sophia.inria.fr/acacia/caus#c287">
          <gal:agent> <gal:Hypoxaemia/> </gal:agent>
          <gal:patient> <gal:CardioEvent/> </gal:patient>
          <gal:setting> <gal:DialysisPatient/> </gal:setting>
          <gal:sem> <gal:trigger/> </gal:sem>
        </gal:CausationSchema>
      </gal:hasForCausationSchema>
  </gal:Abstract>
</rdf:RDF>
```

It is worth noticing that the RDF annotation solely keeps semantic information from the instantiated conceptual schema, and is pruned from all syntactic features. Furthermore, a validation analysis on these semantic annotations is elaborated by domain experts that only retain accurate and relevant ones.

5 Related Work

INSYSE is close to pattern matching methods, that deduce concepts from domain semantic markers and through their contextual analysis; COATIS [8] adopts this approach to extract causality relations. INSYSE is also close to ASIUM [7] and OntoLT [4] that stress the importance of grammatical relations to apprehend the interconnections between concepts. However, these approaches perform a direct pattern matching between syntactic parsing and semantic annotation, without an intermediary fine grained semantic construal. Moreover, ASIUM syntactic information process relies on statistics. INSYSE is a semi-automatic knowledge extraction method, close to the approach proposed in [1].

6 Conclusion and Perspectives

We have presented INSYSE, a semi-automatic text annotation method applied in biomedical domain, aiming at construing causation relations implying genes functions in central nervous system pathologies. INSYSE focuses on the acquisition of causation instantiated conceptual schemas, construed by a set of dedicated unification grammar rules processing a lexicon based on the merging of a terminological extraction with a partial syntactic analysis.

The main contribution of our paper is twofold: first we advocate the processing of a fine grained syntactic analysis, by merging a terminological processing with a shallow syntactic parsing; secondly we favour an accurate syntax-semantic interface through a fine grained semantic construal operated by grammar rules and processed by PATR-II grammatical parser.

A first implementation of INSYSE in Java has just been carried out. We are currently making some adjustments to apply our system to the analysis of the whole corpus of 5000 Medline abstracts. From a linguistic viewpoint, we want to evaluate the accuracy of grammar rules we have built together with the whole linguistic process, by analysing the annotation generation – or none generation

– expected for each causality sentences. From the genomic domain viewpoint, experts should validate the relevancy of the semantic annotations iteratively generated.

As further work, versioning and backward interaction between the stages of INSYSE would be useful for validation and adjustment purposes. Second, even if the causality sentence identification in stage 1 is not the core of our work, we can fairly enhance it with more domain-specific causality markers, and those revealed by the terminological analysis of Nomino [6] on the corpus, may also be useful for this task. Finally, the finalization of our annotation construction will be effective with an ontology concept matching stage, so as to obtain consensual semantic annotations. This stage aims at mapping each term filling our PATR-II conceptual schemas with some GALEN [11] concepts. Two different approaches are currently tested for this mapping task: the first one relies on a lexicographic similarity calculus, based on tokens or lemmas analysis; the second one relies on an ontology integration based method, by using semantic similarity calculus described in [13].

References

1. Aussenac-Gilles, N., Biebow, B., Sulzman, S.: Revisiting Ontology Design: a methodology based on corpus analysis. In *Proceedings of EKAW'2000* (2000) 172-188
2. Bourigault, D., Fabre, C. : Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaires 25* (2000) 131-151
3. Briscoe, T., Carroll, J.: Robust accurate statistical annotation of general text. In *Proceedings of LREC'02* (2002) 1499-1504
4. Buitelaar, P., Olejnik, D., Sintek, M.: A Protege plug-in for ontology extraction from text based on linguistic analysis. In *Proceedings of ESWS'04* (2004)
5. Corby, O., Dieng-Kuntz, R., Faron-Zucker, C.: Querying the semantic web with the Corese search engine. In *Proceedings of ECAI'2004* (2004) 705-709.
6. Dumas, L., Plante, A., Plante, P. : *ALN : Analyseur Linguistique de ALN.* ATO, UQAM (1997)
7. Faure, D., Nédellec, C.: A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *Proceedings of LREC workshop on Adapting lexical and corpus resources to sublanguages and applications* (1998) 5-12
8. Garcia, D.: COATIS: a NLP system to locate expressions of actions connected by causality links. In *Proceedings of EKAW'97* (1997) 347-352
9. Maedche, A., Staab, S.: *Comparing Ontologies: Similarity Measures and a Comparison Study.* Internal Report, University of Karlsruhe (2001)
10. Nuyts, J.: *Aspects of a Cognitive-Pragmatic Theory of Language.* Benjamins (1992)
11. Rector, A., Gangemi, A., Galeazzi, E., Glowinski, A., Rossi-Mori, A.: The GALEN Model Schemata for Anatomy. In *Proceedings of MIE'94* (1994)
12. Shieber, S.M.: *An Introduction to Unification-Based Approaches to Grammar.* CSLI Lecture Notes Series, vol. 4. University of Chicago Press, Chicago (1986)
13. Wang, H., Azuaje, F., Bodenreider, O., Dopazo, J.: Gene Expression Correlation and Gene Ontology-Based Similarity: An Assessment of Quantitative Relationships. In *Proceedings of CIBCB'04* (2004) 25-31

An Approach to Automatic Text Production in Electronic Medical Record Systems

Torbjørn Nordgård[1], Martin Thorsen Ranang[2], and Jostein Ven[3]

[1] Department of Language and Communication Studies,
Norwegian University of Science and Technology, NO-7491 Trondheim, Norway
torbjorn.nordgard@hf.ntnu.no
[2] Department of Computer and Information Science,
Norwegian University of Science and Technology, NO-7491 Trondheim, Norway
martin.thorsen.ranang@idi.ntnu.no
[3] KITH, Sukkerhuset, NO-7489 Trondheim, Norway
jostein.ven@kith.no

Abstract. The paper describes basic properties of a sentence generator which requires minimal input information. Input is a set of unstructured semantic concepts, and the generator produces sentences which are compatible with this set by utilizing information from a statistical language model. Output is filtered by a simple context-free grammar. The system is trained on text from electronic medical records, and it is able to produce well-formed sentences in cases involving simple medication prescriptions and symptom descriptions. Basic complexity aspects of the problem are described, and suggestions for efficient implemented generators which manage to produce sentences within acceptable time limits, despite the complexity of the approach, are presented in the final sections.

1 Introduction

Electronic medical records contain information in non-textual format, for instance tables. When descriptions of the patient are produced, as in discharge summaries, non-textual information is consulted and converted into textual form, albeit that physicians often dictate the information. If we assume that the information sources share semantic properties, linguistically speaking, it is tempting to try to suggest textual realizations of this information, even if the information source is completely unstructured. If successful, such an approach can make parts of the text production easier and less time consuming, even for physicians who are used to dictation equipment.

We present a very simple idea for sentence generation and explore some of its consequences: give an unordered set of concepts as input to the generator, and let it make use of a statistical language model from a relevant subject domain in order to figure out how the concepts should be ordered and perhaps enriched by “function” words so that well-formed sentences emerge. The statistical language model is extremely important because it provides statistically based hypotheses

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1187–1194, 2005.

for "local" word ordering, for instance where a diagnosis word is most likely to occur together with modifying adverbs like "strong", "light", etc. We outline an implemented demonstrator in a later section which is capable of taking minimal and unstructured input and produce candidate sentences, which are filtered through a shallow parser and ranked by a stochastic language model.

The output quality of a sentence generator is highly dependent on the information being presented to it. The more detailed and well-structured semantic and preferably syntactic information the system is given as input, the more adequate sentences will be generated. [1] notes that the poorest results of the systems described there can be traced to the fact that input is too underspecified. The tension between high quality generation output and the problems of providing a rich generation base is well-known in the literature, cf. [2] for a discussion of creation of Minimal Recursion Semantics (MRS) structures from non-linguistic input sources.

This paper does not address questions like content and lexical selection (where the input sources to the generator come from), aggregation (how simple sentences can be arranged as more coherent text), discourse specification (when to use pronouns and definite articles), etc. It is solely concerned with simple surface realization from unstructured semantic concepts which have a fairly straightforward textual realization, and we admit, of course, that a system based on our ideas either has to be surrounded by additional components like those mentioned above, or, alternatively, it must be adapted properly to the user's textual profile.

2 Background and Overview

In experiments with natural language generation from tabular information sources in the project *Mobile Electronic Patient Records* (MOBEL) it was discovered that creation of the information basis prior to rule-based language generation turned out to be a quite comprehensive matter. Since the sentences to be generated were fairly simple, we got a feeling of cracking nuts with a sledgehammer. For this reason we hypothesized that it should be possible to produce relevant sentences on the basis of unordered semantic key words and a statistical language model derived from a text corpus from the relevant medical domain.

The corpus consists of discharge summaries that describe and sum up aspects of the treatment of a patient during a hospital stay. It includes information such as patient history (e.g. previous admissions, operations and diseases in the family), biomedical data (such as blood pressure and temperature), results from tests and investigations performed and information about drug prescription and administration.

We assume that input to the sentence generator can be an unstructured set of semantic items, denoted I_{S}, with referential properties, i.e. items like "left", "leg", "pain", and "patient". I_{S} will be transformed into a priority queue Q_{S} of list representations of I_{S}, Q_{S} being an ordered set of permutations of I_{S}. Each list in Q_{S} might be extended by "spurious" items, borrowing a term from [3] in his description of statistical translation models in [4], e.g. in our context base ("dictionary") forms of auxiliary verbs, determiners, etc. These extended lists contain lexical base forms, and when they are transformed to full forms the system has produced a potential sentence. We take, as in statistical machine translation, that each such potential sentence is assigned a probability by some standard stochastic language model.

3 Components

It is irrelevant for the discussion here how I_{S} is created, so we assume it exists prior to the sentence realization process. To ease readability we provide relevant examples in English when possible. For the same reason we use the same example throughout.

3.1 From Semantic Sets to Semantic Queues

Suppose we have the following input set $I_{\mathrm{S}} = \{$"knee", "left", "pain", "patient"$\}$. The goal sentence is *"the patient has pains in his left knee"*, in which the elements of I_{S} are ordered as $\langle$"patient", "pain", "left", "knee"$\rangle$. This is one of the permutations of I_{S}. The number of permutations of I_{S} is equal to $|I_{\mathrm{S}}|!$ (factorial). Obviously, if the cardinality of I_{S} approaches 10 or more, a naïve permutation strategy is doomed to fail. We denote the set of permutations of I_{S} as $\mathrm{perm}(I_{\mathrm{S}})$.

It is desirable that $\mathrm{perm}(I_{\mathrm{S}})$ is ordered according to some empirically motivated strategy, thus turning it into a priority queue. We assume that the language model helps in this process. The formula

$$\mathrm{argmax}_i \; p(\mathrm{perm}(I_{\mathrm{S}}))$$

where i denotes the index of each permutation in $\mathrm{perm}(I_{\mathrm{S}})$, gives the most probable permutation of I_{S}. The probabilities for the permutations can be derived from a corpus, in our case a corpus of base forms taken from a collection of relevant medical record texts.[1] In order to create the queue, remove the most probable permutation from $\mathrm{perm}(I_{\mathrm{S}})$ and put it in the tail of the priority queue Q_{S} (initially empty) and repeat the process until $\mathrm{perm}(I_{\mathrm{S}})$ is empty. Then we have the desired queue.

Each element $C \in Q_{\mathrm{S}}$ is termed a concept list. A concept list has, intuitively, some of the properties of a MRS structure, cf. [5] for an introduction to MRS. Although an MRS is a recursive feature structure the semantic relations of the MRS are ordered via scope restrictions (QEQs in the terminology of MRS). When doing generation on the basis of MRSs the generator makes use of the grammar to enrich the feature structure with syntactic and morpho-syntactic information, thereby yielding a structure which in the end matches a sentence in the language generated by the grammar.

[1] The corpus also has information about initial and final elements of the sentences. Therefore, the ordering process is able to identify permutations with the most probable heads and tails

We will however try to do generation on the basis of a simpler structure (the concept list C) and expand it with a far less rigorous grammar than required by an MRS (in theory without any grammar at all).

4 "Spurious" Additions

A concept list C (like ⟨"patient", "pain", "left", "knee"⟩) often has to be extended by grammatical items in order to pave the ground for a well-formed sentence, for instance determiners, auxiliary verbs and verbs with limited semantic content, most notably auxiliary verbs used with transitive or predicative interpretations. Other items in this group are certain prepositions and adverbs. When these grammatical words are added to the set of items which can be permuted, the complexity of the approach runs out of control immediately if the number of such items to be inserted is unlimited and insertion positions are unrestricted. We put these matters aside for now, but we will in later sections sketch the basics of decoders which are able to circumvent this search space.

5 From Base Form String to Full Form String

Given a sequence of base forms BF (like ⟨"have", "knee", "left", "patient", "pain"⟩) there are several ways of transforming BF into an ordinary sentence (hidden Markov models (HMMs), finite-state transducers, etc.). Establishing an n-best list seems most appropriate, and the first item in the list which satisfies the grammatical constraints of a shallow parser will be assigned the best rank. This is the strategy in the implemented demonstrator below. Each candidate sentence should ideally be checked by a robust parser, thereby avoiding ungrammatical sentences.

We noted earlier that the operations in this approach are computationally intractable. In this section we sketch the most central properties of two possible generators. Both of them make use of a version of the A* algorithm as used in artificial intelligence, see e.g. [6]. As it turns out they resemble decoders used in e.g. speech processing and HMM tagging, and they will henceforth be called decoders.

5.1 Decoder 1

The crucial aspect of any efficient decoder is its ability to avoid paths which cannot lead to successful analysis results, in our case avoiding to consult the entire search space which the generation problem creates. Given the statistical language model we can for instance choose an initial path by setting priorities on bigrams and trigrams which are attested in the training data. In addition we might prefer sequences without "spurious" words, or avoid sequences with more than a fixed amount of additional words.

The cost function is, as usual in these types of decoders, the central tool which assigns values to hypotheses in the priority queue. This function determines the "price" for extending a word sequence with a new word. We also need a success criterion, and an intuitively relevant success criterion for a potential sentence is that all concepts in I_{S} are present in the sequence. Thus, when the decoder includes a member from I_{S} in the hypothesized base form sequence, no cost is added. Another low cost feature is the existence of a trigram from the rightmost string element to its hypothesized continuation. A somewhat simplified version of the cost function is shown in Algorithm 1.

Algorithm 1 Calculate the cost of extending a word sequence with a new word

```
function COST($word_1, word_2, word_3$)                    ▷ $word_3$ is a new candidate
    if $(word_3 \in inputConcepts) \lor (word_3 \in gramWords)$ then
        if TRIGRAMVERIFICATION($word_1$, $word_2$, $word_3$) then
            $cost \leftarrow trigramCost$
        else if BIGRAMVERIFICATION($word_2$, $word_3$) then
            $cost \leftarrow bigramCost$
        else if UNIGRAMVERIFICATION($word_3$) then
            $cost \leftarrow unigramCost$
        else
            $cost \leftarrow highCost$
        end if
    end if
    if $word_3 \in gramWords$ then
        $cost \leftarrow cost \cdot gramCost$                   ▷ penalize grammatical words
    end if
    return $cost$
end function
```

The functions {TRI,BI,UNI}GRAMVERIFICATION() consult the training data in order to verify whether the new word has appeared in the proposed context. The values for *trigramCost*, *bigramCost*, *unigramCost*, *highCost* and *gramCost* are set globally in order to facilitate easy experimentation with the cost function. They default to 0, 10, 50, 500 and 1000, respectively.

We have chosen to implement a breadth-first version of the decoder. Unfortunately, the search space grows exponentially, especially when many spurious words are allowed. For this reason there is a limit on the size of the priority queue ("beaming"). This approach is quite similar to the decoding strategy described in [7], although their problem domain is different. Finding the optimal limit is an empirical question which partly depends on hardware properties.

The decoder will iterate until the queue contains sequences where all members of I_{S} are included. We assume a constant k which limits the number of spurious words, and guarantees the algorithm to halt. The system will propose a sentence only if it survives grammar filtering (the sentence can be parsed).

The performance is promising. $I_{\mathrm{S}} = \{$"kne", "venstre", "smerte", "pasient"$\}$ (English {"knee", "left", "pain", "patient"}) produces the correct base form sequence *"Pasient ha smerte i venstre kne"* (*"patient has pain in left knee"*) in 1700 milliseconds on an Intel Pentium III 866 MHz PC running SICStus Prolog. Transformation to the correct full form string *"Pasienten har smerter i venstre kne"* is done by the HMM converter instantly.

5.2 Decoder 2

In the efforts to make the generator even more efficient we have tried to impose a reliable control structure to the introduction of new elements into the priority queue, as an alternative to beaming. A notorious problem with beaming is that a correct hypothesis could be thrown away in an early phase because lots of eventually incorrect hypotheses received high scores initially. Suppose we have a deterministic finite-state automaton where the labels are part-of-speech (POS) tags (noun, verb, preposition etc), and also suppose that we have a dictionary where all words in the vocabulary are marked with the correct POS tags. Crucially, assume that the automaton characterizes the legal POS sequences. Since the automaton is deterministic a proposed sequence will always be in an unique state from which there is a limited set of POS transitions to other states.[2] Thus, we are guaranteed that the generator never proposes a sequence of words which has an illegal left context, like ⟨"kne", "venstre", "i"⟩ (⟨"knee", "left", "in"⟩).

The cost function is the same as in the previous decoder. But the search space is much smaller, and the highest ranked sentence from the example above is produced in less than 200 milliseconds on the same machine configuration (as compared to 1700 with the previous decoder). This performance is acceptable.

Unfortunately, the important assumption that the automaton characterizes all legal word sequences is problematic. Writing such a device manually is time consuming, and it becomes more and more difficult to modify the automaton as it increases.

In our experiment we produced the network semi-automatically from a simple context-free grammar, i.e. a grammar with atomic rule symbols only (no unification equations). These grammars are much easier to write, extend and maintain than a finite-state automaton, but converting an arbitrary context-free grammar into a finite-state automaton which generates exactly the same language is known to be impossible. However, for our purposes it is sufficient that the automaton excludes as many impossible paths as possible, but it does not matter much if a small amount of ungrammatical sentences are produced. We still have the possibility of using the same context-free grammar in a parser whose purpose is to accept strings.

6 Related Work

The literature on sentence generation contains many approaches where templates are used, in various forms. Generators based on semantically annotated phrase

[2] If a concept has more than one POS tag, this uniqueness property does not hold in its simplest form. But the proposed sequence and one of the tags will still be unique

structure rules are also quite widespread, in particular within rule-based machine translation, see e.g. [8]. The core engine is generally a chart generator, see [9] for the basic ideas. A statistical generator using templates in dialogue systems is presented in [10].

7 Conclusion and Future Work

We have explored formal and practical issues which arise from a simple idea of sentence realization from an unstructured set of semantic concepts. We have used a standard statistical language model together with a simple context-free grammar as the linguistic "knowledge" of the system. The results are promising, both with respect to linguistic quality and performance.

The linguistic quality depends on corpus relevance for the generation tasks, grammar coverage, quality of finite-state approximation and user adaption.

As mentioned in an early section, a system of this type should be adapted to the user's textual profile. If the system is trained on text samples from previous reports written by the user, it will adapt to the user's writing style and vocabulary (including abbreviations). The more the generation task resembles previous examples, the better it is, as in automatic dictation systems.

In future research we will, among other things, explore the impact of high quality resource grammars (see [11] for details of a Norwegian large-scale HPSG grammar) used as filtering devices, bigger training corpora and how an efficiently implemented generator of this type can be included in an experimental electronic medical record system.

References

1. Langkilde-Geary, I.: An empirical verification of coverage and correctness for a general-purpose sentence generator. In: Second International Natural Language Generation Conference (INLG 2002), New York, USA (2002) pp. 17–24
2. Ven, J.: Use of linguistic and non-linguistic resources and methods in natural language generation. PhD thesis, Norwegian University of Science and Technology, Trondheim, Norway (Forthcoming)
3. Knight, K.: A statistical mt tutorial workbook. Unpublished manuscript, `http://www.isi.edu/natural-language/people/knight.html` (1999)
4. Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **2** (1993) pp. 263–311
5. Copestake, A., Flickinger, D., Sag, I.A.: Minimal recursion semantics: an introduction. Technical report, CSLI, Stanford University, Stanford, CA, USA (1997)
6. Nilsson, N.J.: Artificial Intelligence: A New Synthesis. Morgan Kaufmann, San Francisco, CA, USA (1998)
7. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the 2003 Human Language Technology Conference and Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Edmonton, Canada (2003) pp. 48–54

8. Oepen, S., Dyvik, H., Lønning, J.T., Velldal, E., Beermann, D., Carroll, J., Flickinger, D., Hellan, L., Johannessen, J.B., Meurer, P., Nordgård, T., Rosén, V.: Som å kapp-ete med trollet? towards mrs-based norwegian–english machine translation. In: Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation, Baltimore, MD, USA (2004)
9. Kay, M.: Chart generation. In: 34th Annual Meeting of the Association for Computational Linguistics, University of California, Santa Cruz, California, USA, Morgan Kaufmann Publishers / ACL (1996) pp. 200–204
10. Ratnaparkhi, A.: Trainable approaches to surface natural language generation and their application to conversational dialog systems. Computer, Speech & Language **16** (2002) pp. 435–455
11. Hellan, L., Haugereid, P.: Norsource—an exercise in the matrix grammar building design. In Bender, E., Flickinger, D., Fouvry, F., Siegel, M., eds.: Proceedings of ESSLI Workshop: Ideas and Strategies for Multilingual Grammar Development, Vienna, Austria (2003)

gProt: Annotating Protein Interactions Using Google and Gene Ontology

Rune Sætre[1], Amund Tveit[1,3], Martin Thorsen Ranang[1], Tonje S. Steigedal[2], Liv Thommesen[2], Kamilla Stunes[2], and Astrid Lægreid[2]

[1] Department of Computer and Information Science,
Norwegian University of Science and Technology,
N-7491 Trondheim, Norway
{rune.saetre,amund.tveit,martin.ranang}@idi.ntnu.no
[2] Department of Cancer Research and Molecular Medicine,
Norwegian University of Science and Technology,
N-7491 Trondheim, Norway
{tonje.strommen,liv.thommesen,kamilla.stunes,astrid.laegreid}@ntnu.no
[3] Norwegian Centre for Patient Record Research
Norwegian University of Science and Technology,
N-7491 Trondheim, Norway

Abstract. With the increasing amount of biomedical literature, there is a need for automatic extraction of information to support biomedical researchers. Due to incomplete biomedical information databases, the extraction cannot be done straightforward using dictionaries, so several approaches using contextual rules and machine learning have previously been proposed. Our work is inspired by the previous approaches, but is novel in the sense that it combines Google and Gene Ontology for annotating protein interactions. We got promising empirical results - 57.5% terms as valid GO annotations, and 16.9% protein names in the answers provided by our system gProt. The total error-rate was 25.6% consisting mainly of overly general answers and syntactic errors, but also including semantic errors, other biological entities (than proteins and GO-terms) and false information sources.

Keywords: Biomedical Literature Data Mining, Gene Ontology, Google API

1 Introduction

With the increasing importance of accurate and up-to-date databases about proteins and genes for research, there is a need for efficient ways of updating these databases by extracting information from biomedical research literature [8, 20, 21], e.g. those indexed in MEDLINE. Examples of information resources containing such information are LocusLink, UniGene and Swiss-Prot for protein info and the Gene Ontology for semantic labels.

Due to the large and rapidly growing amounts of biomedical literature, the extraction process needs to be more *automatic* than previously. Current extraction approaches have provided promising results, but they are not sufficiently

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1195–1203, 2005.

accurate and scalable. Methodologically all the suggested approaches belong to the *information extraction field* [3], and in the biomedical domain they range from simple automatic methods to more sophisticated, but slightly more manual, methods. Good examples are: Learning relationships between proteins/genes based on co-occurrences in MEDLINE abstracts (e.g. [9]), *manually* developed information extraction rules (e.g. [22]), information extraction (e.g. protein names) classifiers trained on *manually* annotated training corpora (e.g. [1]), and classifiers trained on *automatically* annotated training corpora [19]).

1.1 Research Hypothesis

Internet Search Engines such as Google, Yahoo and MSN Search are the world's largest readily available information sources, also in the biomedical domain. Based on promising results from recent work on using Google for semantic annotation of biomedical literature [16], we are encouraged to investigate if Google can be used to find protein interactions that match the Gene Ontology (GO). This leads to the hypothesis:

Can Internet Search engines such as Google be used to annotate protein interactions in the Gene Ontology framework?

The rest of this paper is organized as follows. Section 2 describes the materials used, section 3 presents our method, section 4 presents empirical results, section 5 describes related work, section 6 discusses our approach, and last the conclusion and future work.

2 Materials

See fig. 1 for an overview of the system. As input for our experiments we used the following:

- 10 proteins that are already well-known to our biology experts.
- 37 verb-templates suggested by Martin et. al (LexiQuest) [12].

Proteins

The following proteins were used as input to the system.

Proteins used
'EGF', 'TNF', 'CCK', 'gastrin', 'CCKAR', 'CCKBR', 'p53', 'ATF1', 'CREB', 'CREM'.

In addition, each protein is also described by several other names or synonyms in the literature. E.g. gastrin is also known as 'g14', 'g17', 'g34', 'GAS', 'gast', 'gastrin precursor', 'gastrin 14', etc. So our biologists compiled a list of roughly 10 synonyms for each protein, *giving us about 100 terms total to annotate.*

Interaction Verbs

We selected our interaction verb templates from table 1 in [12]. They had a list of 44 verbs, but we chose to use only 37 of these verbs. The reason for this is that we are focusing on simple statements like "gastrin activates ...", with the object of the verb following directly after the verb template. The following table shows the original list of verbs, with the removed ones in parenthesis.

Verb templates used
acetylates, activates, (antagonizes), associates with, (attenuates), (binding to), binds, blocks, (bonds), (complex), deactivates, decreases, degrades, dephosphorylates, dimerizes, dissociates from, downregulates, forms complex with, hydrolyses, inactivates, increases, induces, inhibits, interacts with, links, mediates, (oligomerizes), overexpresses, phosphorylates, potentiates, precipitates with, reacts with, recruits, (reduces), regulates, releases, represses, stimulates, transactivates, transduces, transforms, triggers, ubiquitinates, upregulates,

3 Our Approach

We have taken a modular approach where every submodule can easily be replaced by other similar modules in order to improve the general performance of the system. There are five modules in the system. The first one sets up the search queries, the second runs the queries against Google, the third one tokenizes the results, the fourth parses the tokenized text, and the fifth and last module extracts all the results and presents them to the human evaluators. See figure 1.

1. **Data Selection.** N (=100) protein names are combined with M (=37) verb templates, giving a total of N x M (3700) queries to run against Google.
2. **Google.** The queries are fed to the PyGoogle module which allows 1000 queries to be run against the Google search engine every day with a personal password key. In order to maximize the use of this quota, the results of every query are cached locally, so that each given query will be executed only once. If a search returns more than ten results, the resultset can be expanded by ten at a time, at the cost of one of the 1000 quota-queries every time. We decided to use up to 30 results for each query in this experiment.
3. **Tokenization.** The text is tokenized to split it into meaningful tokens, or “words”. We use a simple WhiteSpaceTokenizer from NLTK, where every special character (like ( ) ” ’ - , and .) is treated as a separate token.
4. **Parsing.** Each returned hit from Google contains a “snippet” with the given query phrase and approximately ten words on each side of it. We use some simple regular grammars to match the phrase and the words following it. If the next word is a noun it is returned. Otherwise, adjectives are skipped until a noun is encountered, or a “miss” is returned.
5. **Expert Evaluation.** The results were merged so that all synonyms were treated as if the main protein name had been used in the original query. Then the results were put into groups (one group for each protein-verb pair) and sorted alphabetically within that group. These results were then presented to the biologists, who evaluated the usefulness of our results from Google.

4 Empirical Results

Fig. 2 and 3 show the results. The first one shows that more than half of the extracted terms were terms that could be used to annotate the given protein

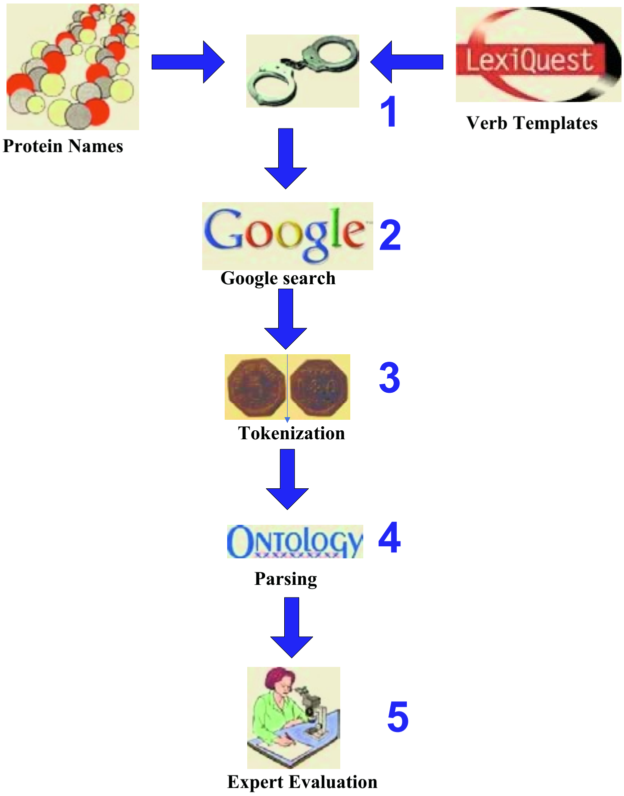


Fig. 1. Overview of Our Approach (named gProt)

according to the Gene Ontology (GO). Around one fifth of the results contained an identifiable protein name that could be stored as a protein-protein interaction. Only one quarter of the terms were deemed not useful. The different kinds of "not useful"-errors can be read out of fig. 3.

5 Related Work

Our specific approach was on using Google and Gene Ontology for annotating protein interactions. We haven't been able to find other work that does this, but the closest are Dingare et al., that uses results from Google search as a feature for a maximum entropy classifier used to detect protein and gene names [5, 6], and our previous work on semantic annotation of proteins (i.e. tagging of individual proteins, not their GO relation) [16]. Google has also been used for semantic tagging outside of the biomedical field, e.g. in Cimiano and Staab's PANKOW system [2] and in [4, 7, 10, 11, 13, 17].

A comprehensive overview of past methods for protein-related information extraction is provided in [18].

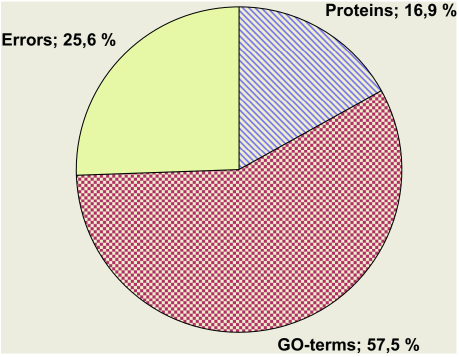


Fig. 2. Main Results

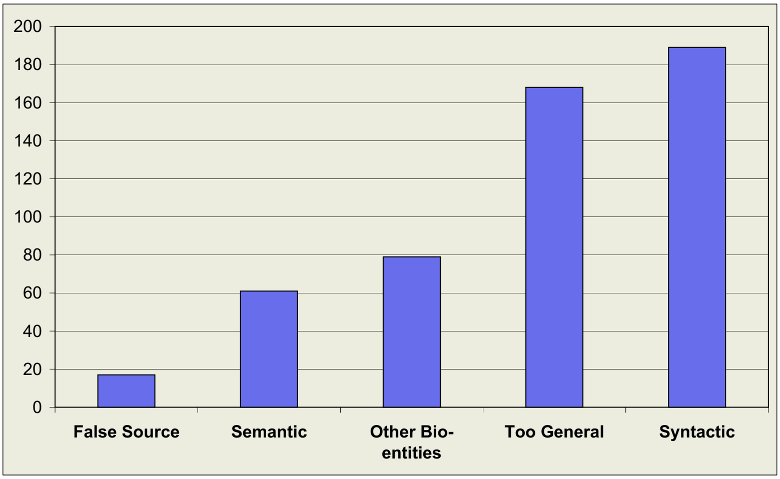


Fig. 3. Breakdown of Errors

6 Discussion

In the following section we discuss our approach step-by-step. (The steps as presented in fig. 1.)

1. **Data Selection.** The results were inspected by cancer researchers, so the focus was naturally on proteins with a role in cancer development, and more specifically cancer in the stomach. One such protein is gastrin, used as a running example in this article. In the experiment we used ten such protein names with around ten synonyms for each. The large number of synonyms used for each original protein name gave us a valuable increase in the recall of expected facts from Google.

2. **Google.** Since we decided to download up to 3 (times 10) results for each query, we had to do around 11.000 queries. This took almost two weeks, because of Google's restraint to only run 1000 queries per day. If we want to scale up this method in the future, we would probably have to pay Google to let us do more queries per day, or consider using the recently announced Yahoo API that allows 5.000 queries per day. The number of returned GO-processes was over 50%, which is very promising for automatic annotation, considering that no information has been used in the process to match GO-terms more often than e.g. protein names.
3. **Tokenization.** Most of the "errors" are syntactic errors, and many of the syntactic errors occur because of bad tokenization, mainly because a lot of the returned words are just parts of multi-word-tokens. Also, many of the words are not nouns at all, so they are not suitable class names in the first place. In the future more work should be done in the tokenization phase. The WhiteSpaceTokenizer was used because it is easy and fast, but with some sort of NP-clustering and parentheses handling, almost half of the errors could be removed. One example of NP clustering is protein names, such as "g-protein coupled receptor (GPCR)".
   How to deal with parentheses? Sometimes they are important parts of a protein name (often part of "formulae" describing the protein), and other times they are just used to state that the words within them aren't that important. And the worst problem is that they are quite often "unbalanced", either because of typing errors, "1) 2) 3)"-style numbering, or smileys.
4. **Parsing.** We used a really simple grammar to extract the interacting terms from what Google returned. It can be summed up as: After the template, keep reading words until a "stop-word" is encountered. As "stop-words" we used some common prepositions, in addition to full-stop punctuation (.,;?!). There is obviously room for a lot of improvements here, e.g. using more advanced Natural Language Understanding techniques.
5. **Expert Evaluation.** The evaluation was quite simple, just focusing on deciding whether this way of using Google to do information extraction is worth pursuing or not. Since the tokenization and grammar modules aren't perfect yet, the biologist also had access to the complete snippets (and the corresponding homepage) in their evaluation work. It is now obvious to us that we should keep developing this system, since almost three out of four results were relevant, and many of them also novel, information.

7 Conclusion and Future Work

This paper presents a novel approach - gProt - using Google to find semantic (GO-) annotations for specific proteins.

We got empirically promising results - 57.5% semantic annotation classes, and 16.9% protein names in the answers provided by gProt. This means that 74.4% of the results are useful. This encourages further work, possibly in combination with other approaches (e.g. rule based information extraction methods), in order to improve the overall accuracy. In the similar task of protein name identification, recently presented precision scores ranges from 70 to 75% [1]. Hopefully, more advanced methods will greatly reduce the number of errors (useless information), which is currently at 25.6%. Disambiguation is another issue that needs to be further investigated, because sometimes different search-results are really just one single identity, because of synonyms and acronyms for example. Other opportunities for future work include:

- Improve tokenization. Just splitting on whitespace and punctuation characters is *not* good enough. In biomedical texts non-alphabetic characters such as brackets and dashes need to be handled better.
- Search for other verb templates using Google. E.g. Which templates give the best results, and what about negations (“does not activate ...”)?
- Investigate whether the Google ranking is correlated with the accuracy of the proposed semantic tag. Are highly ranked pages better sources than lower ranked ones?
- Test our approach on larger datasets, e.g. using *all* the returned results from Google.
- Combine this approach with more advanced natural language parsing techniques in order to improve the accuracy, [14, 15].
- In order to find multiword tokens, one could extend the search query *(“X activates”)* to also include neighboring words of X, and then see how this affects the number of hits returned by Google. If there is no reduction in the number of hits, this means that the words are “always” printed together and are likely constituents in a multiword token. If you have only one actual hit to begin with, the certainty of the previous statement is of course very weak, but with increasing number of hits, the confidence is also growing.
- In this experiment very crude Part Of Speech (POS) tagging is done, so our results can be seen as a baseline for this kind of experiment. In the future we want to improve the results, for example by utilizing better grammars, and more advanced natural language understanding techniques.

Acknowledgements

We would like to thank Waclaw Kusnierczyk and Tore Amble for continuous support.

References

1. Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun Kumar Ramani, and Yuk Wah Wong. Comparative Experiments on Learning Information Extractors for Proteins and their Interactions. *Journal Artificial Intelligence in Medicine: Special Issue on Summarization and Information Extraction from Medical Documents (Forthcoming)*, 2004.
2. Philipp Cimiano and Steffen Staab. Learning by Googling. *SIGKDD Explorations Newsletter*, 6(2):24–34, December 2004.

3. J. Cowie and W. Lehnert. Information Extraction. *Communications of the ACM*, 39(1):80–91, January 1996.
4. Stephen Dill, Nadav Eiron, David Gibson, Daniel Gruhl, R. Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin, and Jason Y. Zien. SemTag and seeker: bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the Twelfth International World Wide Web Conference, WWW2003*, pages 178–186. ACM, 2003.
5. Shipra Dingare, Jenny Finkel, Christopher Manning, Malvina Nissim, and Beatrice Alex. Exploring the Boundaries: Gene and Protein Identification in Biomedical Text. In *Proceedings of the BioCreative Workshop*, March 2004.
6. Shipra Dingare, Jenny Finkel, Christopher Manning, Malvina Nissim, Beatrice Alex, and Claire Grover. Exploring the Boundaries: Gene and Protein Identification in Biomedical Text. Submitted to BMC Bioinformatics, 2004.
7. Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. Submitted to Artificial Intelligence, 2004.
8. Jun ichi Tsuji and Limsoon Wong. Natural Language Processing and Information Extraction in Biology. In *Proceedings of the Pacific Symposium on Biocomputing 2001*, pages 372–373, 2001.
9. Tor-Kristian Jenssen, Astrid Lægreid, Jan Komorowski, and Eivind Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–28, May 2001.
10. Vinay Kakade and Madhura Sharangpani. Improving the Precision of Web Search for Medical Domain using Automatic Query Expansion. Online, 2004.
11. Udo Kruschwitz. Automatically Acquired Domain Knowledge for ad hoc Search: Evaluation Results. In *Proceedings of the 2003 Intl. Conf. on Natural Language Processing and Knowledge Engineering (NLP-KE'03)*. IEEE, 2003.
12. Eric P. G. Martin, Eric G. Bremer, Marie-Claude Guerin, Catherine DeSesa, and Olivier Jouve. Analysis of Protein/Protein Interactions Through Biomedical Literature: Text Mining of Abstracts vs. Text Mining of Full Text Articles. In *Proceedings of the Knowledge Exploration in Life Science Informatics (KELSI2004) Symposium*, volume 3303 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 96–108. Springer-Verlag Heidelberg, 2004.
13. David Parry. A fuzzy ontology for medical document retrieval. In *Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation - Volume 32*, pages 121–126. ACM Press, 2004.
14. Rune Sætre. GeneTUC, A Biolinguistic Project. (Master Project) Norwegian University of Science and Technology, Norway, June 2002.
15. Rune Sætre. Natural Language Processing of Gene Information. Master's thesis, Norwegian University of Science and Technology, Norway and CIS/LMU Munchen, Germany, April 2003.
16. Rune Sætre, Amund Tveit, Tonje Strœmmen Steigedal, and Astrid Lægreid. Semantic Annotation of Biomedical Literature using Google. In Dr. Marina Gavrilova, Dr. Youngsong Mun, Dr. David Taniar, Dr. Osvaldo Gervasi, Dr. Kenneth Tan, and Dr. Vipin Kumar, editors, *Proceedings of the International Workshop on Data Mining and Bioinformatics (DMBIO2005)*, Lecture Notes in Computer Science (LNCS) (Forthcoming), Singapore, May 2005. Springer-Verlag Heidelberg.
17. Urvi Shah, Tim Finin, and Anupam Joshi. Information Retrieval on the Semantic Web. In *Proceedings of CIKM 2002*, pages 461–468. ACM Press, 2002.

18. Hagit Shatkay and Ronen Feldman. Mining the Biomedical Literature in the Genomic Era: An Overview. *Journal of Computational Biology*, 10(6):821–855, 2003.
19. Amund Tveit, Rune Sætre, Tonje S. Steigedal, and Astrid Lægreid. ProtChew: Automatic Extraction of Protein Names from Biomedical Literature. In *Proceedings of the International Workshop on Biomedical Data Engineering (BMDE 2005, in conjunction with ICDE 2005)*, Tokyo, Japan, April 2005. IEEE Press (Forthcoming).
20. Limsoon Wong. A Protein Interaction Extraction System. In *Proceedings of the Pacific Symposium on Biocomputing 2001*, pages 520–530, 2001.
21. Limsoon Wong. Gaps in Text-based Knowledge Discovery for Biology. *Drug Discovery Today*, 7(17):897–898, September 2002.
22. Hong Yu, Vasileios Hatzivassiloglou, Carol Friedman, Andrey Rzhetsky, and W. John Wilbur. Automatic Extraction of Gene and Protein Synonyms from MEDLINE and Journal Articles. In *Proceedings of the AMIA Symposium 2002*, pages 919–923, 2002.

Physiological Modeling and Simulation for Aerobic Circulation with Beat-by-Beat Hemodynamics

Kenichi Asami

Faculty of Engineering, Kyushu Institute of Technology,
1-1, Sensui, Tobata, Kitakyushu 804-8550, Japan
asami@mns.kyutech.ac.jp

Abstract. In this paper a physiological modeling and simulation system for aerobic circulation in fitness exercise is presented, where the cardiovascular system model predicts pulsatile blood flows in response to exercising levels. Designing a circulatory monitor and simulator to allow interaction of human with health enhancement devices enables to develop intelligent fitness machines. The modeling and simulation system for the human circulation model is vital to realize the safe and effective machines. In order to construct the better architecture of assisting framework, the modeling and simulation system for human circulation model previous to fitness work could contribute to make the effective schemes for rehabilitation, dieting, health enhancement, or physical strengthening.

1 Introduction

A lot of physiological models aim to represent various movement and disease quantitatively. Important parameters can be adjusted in order to predict physical responses to perturbations of system parameters and to confirm causal relationships between parameters and variables in the model. In this paper, a physiological modeling and simulation system to compile human circulation model for being connected to a fitness machine and the experiment interface is proposed. The basic human circulation model is integrated with cardiovascular system and exercise control models. The modeling and simulation system enables to modify the basic model in accordance with simple personal parameters related to athletic ability. Namely, the physiological modeling system enables to access to the basic human circulation model in order to manipulate model parameters, to combine physiological subsystems, and to modify mathematical representations of the model.

The basic human model includes principal physiological subsystems from the respiratory to the hormonal system. Moreover, the cardiovascular system model emulating beat-by-beat heart motion to analyze pulsatile blood flow and the exercise control model evaluating athletic ability according to personal parameters are integrated with the human circulation model. The integrative physiological model enables to predict hemodynamic conditions in fitness training in order to indicate the design concept for health enhancement machines.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1204–1209, 2005.

2 Modeling Integrative Circulatory System

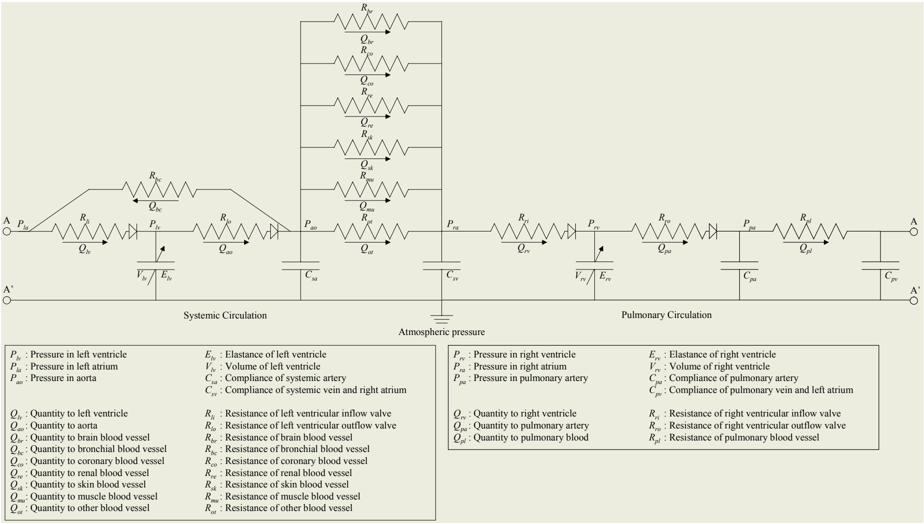
The human circulation model is described on the basis of a representative one called Human developed by Coleman and Randall [1], [2]. The Human model consists of 25 physiological subsystems of principal organs and functions with 321 variables, 70 parameters, and 707 lines of calculation. The 25 subsystems implement HEART (heart/reflex interaction), CARDFUNC (cardiac function), REFLEX-1 (sympathetic nerve reflexes), REFLEX-2 (vagus nerve reflexes), CIRC (pulmonary circulation), O_2 (oxygen balance), CO_2 (carbon dioxide balance), VENT (control of ventilation), GAS (gas exchange), HORMONES (basic renal hormones), KIDNEY (status of kidneys), RENEX (renal excretion), FLUIDS (fluid infusion and loss), WATER (water balance), NA (sodium balance), ACID/BASE (acid/base balance), UREA (urea balance), K (potassium balance), PROTEIN (protein balance), VOLUMES (volume distribution), BLOOD (blood volume and red cell mass), TEMP (temperature regulation), EXER (control of exercise), HEMOD (hemodialysis), and DRUGS (pharmacology) as modules.

The Human model as a basic model, however, does not contain the concept of hemodynamics to analyze pulsatile blood flow. Therefore the cardiovascular system model representing beat-by-beat hemodynamics is connected to the Human model into the integrative physiological subsystems in the whole human body.

2.1 Cardiovascular System Model

The monitoring and diagnosis of circulation are important for adjusting exercise levels in fitness training. The cardiovascular system model describes blood flow in a circulation loop based on ventricular elastance as cardiac function. The time-varying ventricular elastance proposed by Suga defines properties of the heart during cardiac cycle [3]. The definition of the time-varying ventricular elastance as the ratio of ventricular pressure to volume indicates, $E_v(t) = P_v(t)/(V_v(t)-V_0)$ where the inferior $_v$ refers to the ventricle and V_0 represents the unstressed volume of ventricles. During the systolic phase of the ventricle, elastance rises rapidly, and the rise ceases at ejection. During the diastolic phase of the ventricle, elastance falls rapidly in isovolumetric relaxation, and is almost constant in passive filling.

We made the cardiovascular system model shown in Fig. 1, which is consisting of systemic and pulmonary circulations [4]. In this model, the variable capacitor predefined as the ventricular elastance shows heart activity that repeats contraction and relaxation. The fixed compliance represents the vascular suppleness and the blood reservoir. The diode represents the ventricular valve for inflow and outflow of the blood chamber. The aortic blood flow from the left ventricle (Q_{ao}) branches into the brain (Q_{br}), coronary (Q_{co}), kidney (Q_{re}), skin (Q_{sk}), muscle (Q_{mu}), and other vasculature (Q_{ot}), and it joins in the vena cava and the right atrium (Q_{rv}). The about 5% of left ventricular output flows to bronchi (Q_{bc}) and then flows back to left atrium. The systemic circulation can be described by simultaneous differential equations of blood flow in the left ventricle, the aorta, and the vena cava as follows:



R_{br}
Q_{br}
R_{co}
Q_{co}
R_{re}
Q_{re}
R_{sk}
Q_{sk}
R_{mu}
Q_{mu}
R_{bc}
Q_{bc}
A
P_{la}
R_{li}
Q_{lv}
P_{lv}
R_{lo}
Q_{ao}
P_{ao}
R_{ot}
Q_{ot}
P_{ra}
R_{ri}
Q_{rv}
P_{rv}
R_{ro}
Q_{pa}
P_{pa}
R_{pl}
Q_{pl}
A
V_{lv}
E_{lv}
C_{sa}
C_{sv}
V_{rv}
E_{rv}
C_{pa}
C_{pv}
A'
A'
Systemic Circulation
Pulmonary Circulation
Atmospheric pressure
P_{lv} : Pressure in left ventricle
P_{la} : Pressure in left atrium
P_{ao} : Pressure in aorta
E_{lv} : Elastance of left ventricle
V_{lv} : Volume of left ventricle
C_{sa} : Compliance of systemic artery
C_{sv} : Compliance of systemic vein and right atrium
Q_{lv} : Quantity to left ventricle
Q_{ao} : Quantity to aorta
Q_{br} : Quantity to brain blood vessel
Q_{bc} : Quantity to bronchial blood vessel
Q_{co} : Quantity to coronary blood vessel
Q_{re} : Quantity to renal blood vessel
Q_{sk} : Quantity to skin blood vessel
Q_{mu}: Quantity to muscle blood vessel
Q_{ot} : Quantity to other blood vessel
R_{li} : Resistance of left ventricular inflow valve
R_{lo} : Resistance of left ventricular outflow valve
R_{br} : Resistance of brain blood vessel
R_{bc} : Resistance of bronchial blood vessel
R_{co} : Resistance of coronary blood vessel
R_{re} : Resistance of renal blood vessel
R_{sk} : Resistance of skin blood vessel
R_{mu}: Resistance of muscle blood vessel
R_{ot} : Resistance of other blood vessel
P_{rv} : Pressure in right ventricle
P_{ra} : Pressure in right atrium
P_{pa} : Pressure in pulmonary artery
E_{rv} : Elastance of right ventricle
V_{rv} : Volume of right ventricle
C_{pa}: Compliance of pulmonary artery
C_{pv}: Compliance of pulmonary vein and left atrium
Q_{rv} : Quantity to right ventricle
Q_{pa}: Quantity to pulmonary artery
Q_{pl} : Quantity to pulmonary blood
R_{ri} : Resistance of right ventricular inflow valve
R_{ro} : Resistance of right ventricular outflow valve
R_{pl} : Resistance of pulmonary blood vessel

Fig. 1. The cardiovascular system model consists of systemic and pulmonary circulations

$$\frac{dV_{lv}}{dt} = Q_{lv} - Q_{ao} \ . \tag{1}$$

$$\frac{dP_{ao}}{dt} = \frac{Q_{ao} - Q_{br} - Q_{co} - Q_{re} - Q_{sk} - Q_{mu} - Q_{ot} - Q_{bc}}{C_{sa}} \ . \tag{2}$$

$$\frac{dP_{ra}}{dt} = \frac{Q_{br} + Q_{co} + Q_{re} + Q_{sk} + Q_{mu} + Q_{ot} - Q_{rv}}{C_{sv}} \ . \tag{3}$$

The blood flow to individual parts in the systemic circulation can be also determined from the relationship between pressure and resistance. For instance, the blood flow to a muscle blood vessel is calculated as follows:

$$Q_{mu} = \frac{P_{ao} - P_{ra}}{R_{mu}} \ . \tag{4}$$

By the same method for the pulmonary circulation, the output of the right ventricle (Q_{pa}) flows to the lungs (Q_{pl}), and it returns to the left atrium (Q_{lv}). The systemic and pulmonary systems are joined in the loop, where the output of the right ventricle passes the lungs and reaches the pulmonary vein and the left atrium. The pulmonary system can be described as simultaneous differential equations of blood flow in the right ventricle, the pulmonary artery, and the pulmonary vein as follows:

$$\frac{dV_{rv}}{dt} = Q_{rv} - Q_{pa} \ . \tag{5}$$

$$\frac{dP_{pa}}{dt} = \frac{Q_{pa} - Q_{pl}}{C_{pa}} \ . \tag{6}$$

$$\frac{dP_{la}}{dt} = \frac{Q_{pl} + Q_{bc} - Q_{lv}}{C_{pv}} \ . \tag{7}$$

The blood flow to the lungs can be also calculated as follows:

$$Q_{pl} = \frac{P_{pa} - P_{la}}{R_{pl}} \ . \tag{8}$$

The cardiovascular system model is connected to the Human model with common variables of vascular resistance, heart rate, and body weight, and products pulsatile blood flow and pressure as the output.

2.2 Exercise Control Model

The responses of respiration, venous contraction, and muscle metabolism for exercise are presented in the Human model. The exercise is defined as the addition of oxygen in blood from 0 to 10,000 ml/min for normal oxygen use 250 ml/min. If the exercise is given, respiration rate, sympathetic activity, venous pressure, and muscular metabolism increase according to the exercise levels. Consequently, cardiac output and venous return rise in the circulatory system. In addition to the functions, the exercise control model is constructed by introducing personal parameters of body weight, height, age, and sex and evaluation variables of maximum oxygen uptake, basal metabolic rate, and body fat percentage, related to fitness exercise.

VO_2max as exercise intensity promotes individual endurance and performance. Understanding personal *VO_2max* in ml/kg/min or aerobic power is the key for enhancing his/her maximum uptake of oxygen, because it indicates the maximum amount of oxygen he/she can take in and utilize concerning the termination of exercise. *VO_2max* is described as the following equation by Wolthuis depending on the age, gender, and fitness habits, where the first coefficient is set to 50.6 for the active level, 45.8 for the moderate level and 43.2 for the sedentary level. For women, the value becomes 75% regardless of age. Thus the maximum oxygen uptake can be assessed by multiplying *VO_2max* by body weight.

$$VO_2\,max = 50.6 - 0.17 \times AGE \ . \tag{9}$$

Basal metabolic rate (*BMR*) is an estimate of how many calories his/her body would burn if he/she was to do nothing but rest for 24 hours. It represents the minimal amount of caloric requirement needed to sustain life including heart beating, lungs breathing, and body temperature normal in a resting individual. The purposes of the fitness training would be health and weight management in many cases. Therefore, *BMR* is an essential index in the exercise control model, and influences to calorie production presented to the user. *BMR* is calculated by the Harris-Benedict equation from weight in kilograms, height in centimeters, and age in years, where the upper one is used for men and the lower for women.

$$\begin{cases} BMR = 66 + (13.75 \times WEIGHT) + (5.0 \times HEIGHT) - (6.76 \times AGE) \\ BMR = 655 + (9.56 \times WEIGHT) + (1.85 \times HEIGHT) - (4.68 \times AGE) \end{cases} . \tag{10}$$

Body composition and health are affected by the amount of body fat because muscle tissue is more compact than fat. Measuring changes in body fat percentage, rather than just measuring changes in weight, can be very motivational for dieting. Body fat percentage is measured by several methods, such as bioelectrical impedance, skin fold measurement, hydrostatic weighing, and infrared interactance. In the exercise control model, body fat percentage as input influences to muscle mobilizing rate in the fitness training.

3 Results

The parameter of gradual exercise intensity of 100W, 200W, and 300W was introduced to the human circulation model. Personal parameters with body weight of 60kg, height of 175cm, age of 40 years old, the male sex, body fat percentage of 20%, and fitness habit of the moderate level were set. 2 hours of a continuous exercise and subsequent 1 hour of a steady state were given to the model.

Fig. 2 shows the simulation results for physiological variables related to exercise evaluation. Lactic acid as an index of exercise intensity is produced by anaerobic metabolism mainly from muscles, and the production is proportional to oxygen debt. In Fig. 2, lactate mass rapidly went up to 250 mmol when exercise intensity was set to 300W from 200W, since it is used as energy to a certain amount of exercise intensity. Muscle O_2 use was proportional to exercise intensity, where the marginal exercise intensity could be determined by being kept muscle O_2 use less than *VO_2max*.

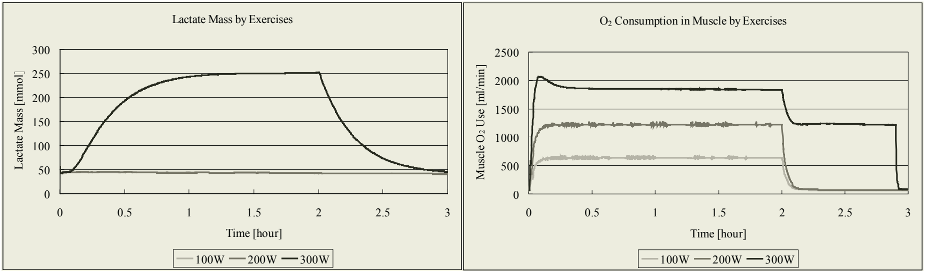


Fig. 2. The simulation results of lactate mass and muscle O_2 use in aerobic circulation

Fig. 3 shows the simulation results for blood flow to muscle and skin in an aerobic state of 5 seconds by the cardiovascular system model according to exercise intensities of 100W, 200W, and 300W. The muscle blood flow (Q_{mu}) was 5649 ml/min for 100W, 9648 ml/min for 200W, and 12520 ml/min for 300W. The skin blood flow (Q_{sk}) was 849 ml/min for 100W, 1063 ml/min for 200W, and 1509 ml/min for 300W. In a resting state, about 25% of cardiac output flows to muscle and skin. In an exercising state, about 85% of cardiac output flows to muscle and skin. In Fig. 3, 62% of cardiac output for 100W, 74% of cardiac output for 200W, and 82% for cardiac output for 300W flowed to muscle and skin in the cardiovascular system model.

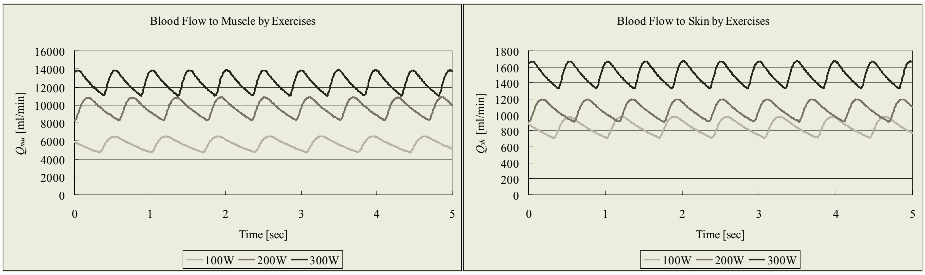


Fig. 3. The simulation results of blood flow to muscle and skin in aerobic circulation

4 Conclusions

This paper describes an integrative physiological modeling and simulation system and the possibility of its application to fitness exercise. Initial simulation using the human circulation model shows the proper correlation with typical exercise physiology. Hemodynamic predictions are also consistent with physiological information, although Coleman's Human model is inadequate in that respect.

For the real application of the aerobic circulation model with the practical use of fitness exercise, the technological development to measure more exact physiological data would be need. Detailed dynamic data for O_2/CO_2 and lactic acid concentration in blood are essential to the sophistication of the exercise control model.

Acknowledgement

This research was supported in part by a Grant-in-Aid for Young Scientists (B) (KAKENHI 16700363) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Coleman, T.G.: A Mathematical Model for the Human Body in Health, Disease, and during Treatment, ISA Trans., Vol. 18, No. 3 (1979) 65–73
2. Randall, J.E.: Microcomputers and Physiological Simulation, Raven, New York (1987)
3. Suga, H.: Time Course of Left Ventricular Pressure-Volume Relationship under Various End-diastolic Volumes, Jap. Heart J., Vol. 10, No. 6 (1969) 509–515
4. Asami, K. and Kitamura, T.: Physiological Simulation by Integrating a Circulatory System Model with Beat-by-beat Hemodynamics, Proc. of International Conference on Knowledge-Based Intelligent Information and Engineering Systems, Part II, (2003) 388–393

Collaborative and Immersive Medical Education in a Virtual Workbench Environment

Yoo-Joo Choi[1], Soo-Mi Choi[2], Seon-Min Rhee[1], and Myoung-Hee Kim[1,3,⋆]

[1] Department of Computer Science and Engineering
Ewha Womans University, Seoul, Korea
{choirina,blue}@ewhain.net

[2] School of Computer Engineering, Sejong University, Seoul, Korea
smchoi@sejong.ac.kr

[3] Center for Computer Graphics and Virtual Reality
Ewha Womans Univesity, Seoul, Korea
Tel.:+82-2-3277-4418, Fax:+82-2-3277-4409
mhkim@ewha.ac.kr

Abstract. In this paper we introduce a framework for an advanced medical education in a table-type VR environment, which is based on an efficient deformable modeling for the interactive manipulation of soft tissues and a realistic motion simulation for dynamic human organs. The proposed framework includes two data storages with 3D/4D models and the 2D anatomic information, and various functional modules that are used to effectively explain human organs by visualizing and synchronizing with the components in the data storages. In our VR environment, a table-type virtual workbench is effectively used for the collaborative and realistic presentation to a group of learners and the video-based tracking environment provides the convenience for a user to select and to manipulate a subject of learning.

1 Introduction

Computer-based medical education has become necessary and indispensable because it permits repetitive practice and non-invasive training based on medical images without the need for cadavers or plastic models. In traditional medical education, invisible human structures were depicted as 2D images in the form of sectioned images through textbooks [1, 2]. Most anatomy textbooks are composed of thick volumes since they include many pictures to convey accurate anatomic details. The National Library of Medicine (NLM) has completed the Visible Human project, which provides images from CT, MRI, and photo-tomography of a cadaver for researchers to develop educational tools and other simulators [3]. Some researchers have established 3D anatomical atlases of the human body with this new resource [4] while others have presented educational systems for medical students based on 2D tomography image [5]. By using the 3D anatomical atlases, medical students can observe internal structures from any viewing direction. However, since static anatomical atlases can not represent the motion of

[⋆] Corresponding author

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1210–1217, 2005.

dynamic organs and the deformation of soft tissues through user interaction, static atlases are insufficient when we want to understand dynamic organs such as a beating heart or the features of soft tissues. Various physically-based modeling methods have been researched in order to represent the deformation of soft tissues. Surgical training systems to focus on the specific organs have been introduced based on these deformation methods[6–8]. However, most of these training systems have used the computer monitor or head-mounted display devices which are not proper for a group training.

In this paper we propose a framework for advanced medical education in a virtual workbench environment, which provides a realistic motion simulation for the dynamic heart and an efficient deformable modeling for the interactive manipulation of soft tissue organs. In our VR environment, several users can also clearly grasp the shape features of human organs during the training through the use of a wide table-type semi-immersive VR workbench and also conveniently use multi-modal interaction through vision-based tracking. Vision-based tracking tracks the movement of the head and the hand without troublesome device cables.

The rest of the paper is organized as follows. We give a brief summary of the features of table-type VR environment and components of the proposed medical education system in Section 2. In Section 3, we describe an efficient deformation technique for the interactive manipulation and a realistic motion simulation of a dynamic organ. Section 4 presents the implementation results of the proposed framework and concludes the paper.

2 An Interactive Virtual Workbench for Collaboration

In this section, we summarize the features of a table-type virtual workbench and depict the main components of the proposed medical education system which is implemented on the table-type VR environment.

The virtual workbench has several characteristics which make it especially efficient for medical applications. First, it is a horizontal-type display environment similar to a real-life operation table. This makes it possible to simulate an operation in a natural way while other VR environoments which usually have a vertical plane as a screen, are unnatural in their interaction. Moreover, since the gradient of the screen can be tilted up to 45°, the user can see various viewpoints of the 3D model and easily manipulate them. Secondly, it provides an efficient workspace for collaborative tasks since the screen of the virtual workbench is wider than that of a monitor-based system or HMD (head mounted display). Several users can stand around the virtual workbench and carry out discussions on the displayed 3D model similar to discussions carried out on the table-type workspaces in the real world. We can extend this environment with two or more virtual workbenches through the network and it can be used for remote collaboration tasks. Lastly, since it provides the feeling of semi-immersion, users can understand the shape features of the subject of interest more clearly with face-to-face communication.

Fig. 1 depicts the hardware setup of our system. The vision tracking system detects the position and orientation of the user and the interaction stick. We attach infrared-reflective luminous balls to the shutter glass for the main user and the interaction stick. We also place infrared lamps near the cameras to distribute infrared light. As we attach

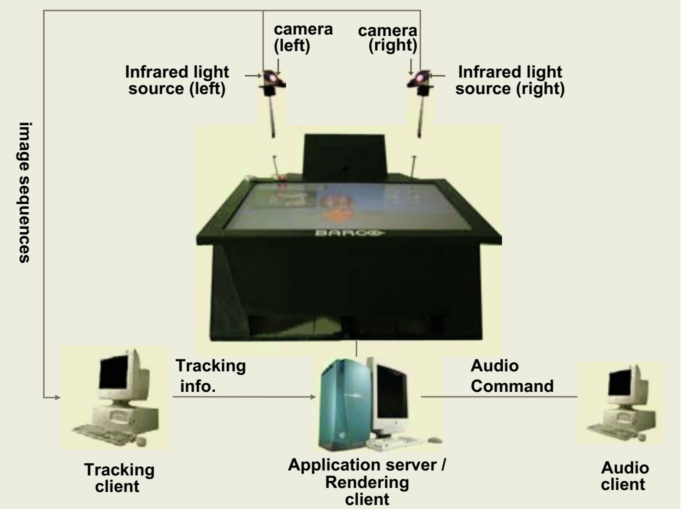


Fig. 1. Hardware system configuration

longpass filters on cameras, only infrared reflective luminous balls appear in the image sequences. The tracking system detects and computes the 3D coordinates of the luminous balls. The viewpoint is changed according to the position of the main user, i.e., the position of the luminous ball on the shutter glass worn by the main user. The main user can select and manipulate a virtual organ through a interaction stick. Fig. 2 shows tracking and interaction devices used in our system. A rendering client is responsible for the interactive rendering of the overall VR education space including 3D/4D models, a 2D information panel and function buttons. An audio client creates the corresponding audio explanation for a selected organ based on the voice processing engine.

The proposed medical education system consists of two data storages including 3D/4D models, 2D anatomic information and various functional modules that are used to effectively explain the human organs by visualizing and synchronizing the components in the data storages as shown in Fig. 3. 3D models in the data storage are built from the Visible-Human data set of NLM [3] and the 4D cardiac model is reconstructed from a motion simulation method as explained in Sec. 3.2. The interactive manipulation not only handles the rigid transformation such as translation and rotation, but also the deformation of the 3D models for the soft tissue organs through the multi-resolutional

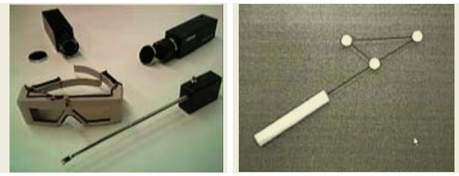


Fig. 2. Tracking and interaction devices. Left: Cameras, shutter glass and calibration stick, Right: Interaction stick

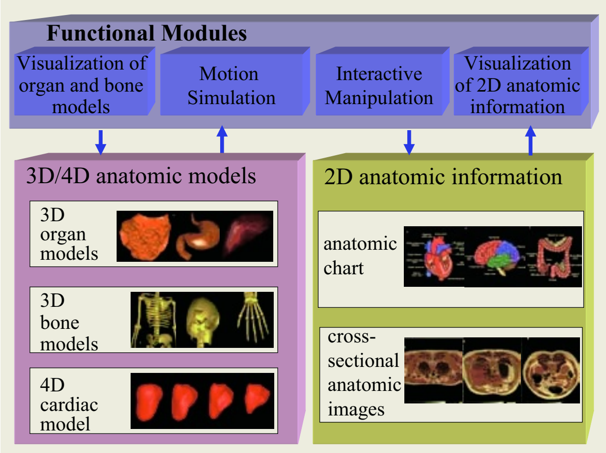


Fig. 3. Main components of the proposed medical education system

approach as explained in Sec. 3.1. A 2D anatomic chart that depicts the details of a corresponding internal organ is displayed on a 2D panel when the user selects the organ in the 3D/4D model space. A 2D cross-sectional anatomic image is also displayed according to the moving plane which is vertically translated onto the whole body through user interaction.

3 Interactive Deformation and Dynamic Simulation of Internal Organs for Immersive Education

In this section, we present a method to simulate the motion of a dynamic heart and interactive deformation of soft tissues for surgery simulation. Motion simulation and interactive deformation can provide an efficient and realistic method for advanced medical education.

3.1 Interactive Deformation of Soft Tissues

In this paper, a spatial adaptation approach of deformable model with physical properties adjustment scheme is applied for the visual realism and simulation efficiency of deformation of soft tissues. Moreover, a progressive force computation scheme using dynamic active component set is presented in order to reduce the total computation cost for solving the numerical system.

The main idea of spatial adaptation is based on the dynamic combination of different detail levels of mass-spring models according to the parts that are subjected to excessive regional external forces or has large curvature variation in order to reduce the number of mass nodes and springs for computation. In the off-line pre-processing phase, the multi-level mesh structure is constructed based on the modified-butterfly subdivision scheme[9]. In the on-line animation phase, the criteria for refinement and simplification are checked for each node. If the criteria are satisfied, the neighboring faces of the node are subdivided or simplified. In the case of simplification, the force and velocity for the refined mid-point node are respectively computed by averaging those of the adjacent nodes on the corresponding edge. In order to maintain the total mass of object and to reduce the numerical instability, we calculate the mass per node for each detail level in the pre-processing step and redistribute the mass nodes in a refined or simplified region only at the time of refinement or simplification, and maintain the value during simulation. The stiffness and damping factors are adjusted by reflecting the level of detail of current spring and masses of nodes incident on the spring in order to maintain the acceleration of a node between different detail levels as Eq. 1. The values for each spring are preserved in the online animation phase.

$$\begin{aligned} KS_l &= 2^l \times KS_0 \times \frac{\sum_{v \in P_l} mass(v)/N_l}{\sum_{v \in P_0} mass(v)/N_0} \\ KD_l &= KD_0 \times \frac{KS_l}{KS_0} \end{aligned} \tag{1}$$

where, KS and KD are the stiffness and damping coefficient for the spring, respectively. N is the number of mass nodes and P means the triangle set in each detail level. l and 0 are the current refined level and the initial mesh level, respectively.

In the case of mass-spring model, the force and velocity are propagated from the nodes affected by external forces over time. Therefore, all springs and nodes have no need to be always included in the force and velocity computation. Therefore, in this study, an active component set is managed based on current force and velocity values of nodes for force computational efficiency. If the current force and velocity for each node are greater than the criterion error, the node are activated, and if either or both of two component nodes of each spring are activated, the spring is activated for spring force computation. Only activated nodes are included in the numerical system for computation of new position and velocity.

In order to validate the representation details and effectiveness of the proposed deformation approach, first, behavioral similarity among models of different detail levels was proved by visual comparison and displacement comparison of the forced nodes over time. Fig. 4 shows the very similar results of deformation of the three models, i.e., coarse, overall refined, and the proposed adaptive refined models, after deformation for 2.3 seconds. As a result of displacement comparison, the displacement of the forced nodes for the refined and proposed adaptive models coincided. On the other hand, the proposed adaptive scheme reduced computational time about 62 % compared to the wholly-refined model. Fig. 5 shows propagation of active springs and nodes over time. Only black nodes and springs in Fig. 5 are included in the numerical system at each simulation step. As a result of progressive force computation using dynamic management of active component set, simulation performance was improved by about 29% in adaptive stomach model and the deformation results of before and after applying active component set were completely identical.

3.2 Motion Simulation of Dynamic Organs

We start with a set of 3D points at each time step of a cardiac cycle in order to simulate the motion of the heart. These points are obtained from the surface of the heart

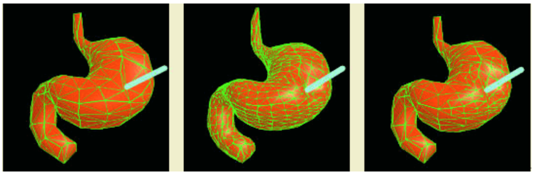


Fig. 4. Comparison of overall behavior according to the same external force in three different models.Left: Coarse model with 188 nodes and 558 springs, Middle: Refined model with 746 nodes and 2232 springs, Right: Proposed adaptive model with 188∼240 nodes and 558∼700 springs

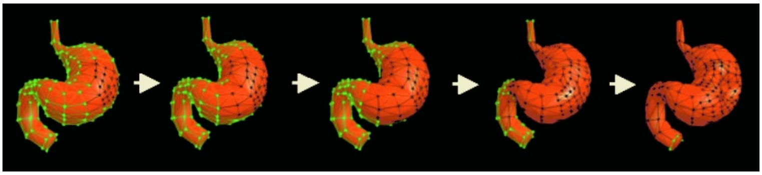


Fig. 5. Propagation of active components over time. Black circles and lines depict activated nodes and springs

which has already been segmented. We decompose the overall motion of the ventricle into global and local motion. In our approach, finite element(FE) analysis is applied to determine the dynamic equilibrium shape of an elastic body. At each time step for the motion tracking, our algorithm finds a translation and rotation that define the global position of the ventricle. The model then deforms itself under forces exerted by virtual springs which connect the model and feature points at the next time step. We use a single 3D blob element which is created with a priori knowledge about the anatomical structure of the ventricle. The blob element is based on superellipsoid with the same size and orientation as the initial reference shape. It is triangulated with the desired geometric resolution, and then the vertices of triangles become FE nodes. We have introduced a new type of Galerkin interpolant, based on a 3D Gaussian, that allows us to solve motion tracking problems efficiently, and to compute robust canonical descriptions for data in 3D space.

The steps for the motion tracking can be summarized as follows. We first create a superellipsoid model for the initial set of 3D points, centered at the center of gravity and rotated on to the principal axes. FE nodes are superimposed on this model, which unifies the geometric and the physical information. The correspondence between 3D points and modal points is established by finding pairs which are bi-directionally closest. Then, mass, stiffness, and mode shape vectors are derived for the FE computation. In general, the number of feature points will be greater than the number of modal displacements to be estimated; thus, we can solve them using a weighted least-squares method. The deformed model is then computed by applying the calculated displacements to the nodal points. At the next time step, this deformed model is translated and rotated to match the center of gravity and principal axes. After which it becomes the undeformed model at the next time step. The positions of the FE nodes are also updated using the calculated nodal displacements. We can extract displacements, speeds, and accelerations from this 4D model. Fig. 6 shows screenshots of the result of motion simulation.

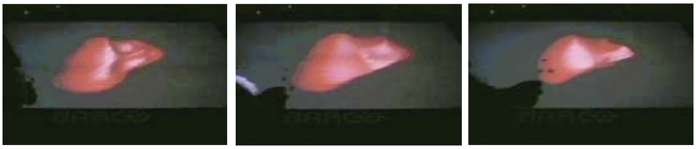


Fig. 6. Motion simulation of the heart

4 Results and Conclusions

In the proposed medical education system, users can select a human organ of interest using a wireless interaction stick. The detailed anatomic information for the selected organ is displayed on the 2D information panel. Fig. 7(a) shows a screenshot when an user select the intestine. When the heart is especially selected, a 3D pumping heart is dynamically rendered as shown in Fig. 6. Fig. 7(b)shows a 2D cross-sectional anatomic image displayed according to the moving plane which is vertically translated onto the whole body through user interaction. Fig. 7(c) shows a group of students being immersed in the study of the human internal organs with the aid of shutter glasses. In this system, the user can examine the internal organs from any viewing point by just moving the main user's head.

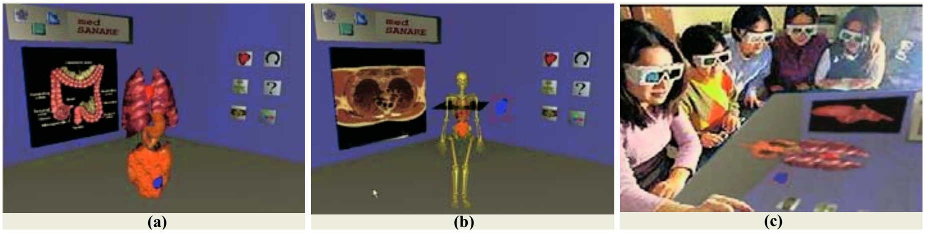


Fig. 7. Screenshots of medical education in VR. (a) 3D anatomic models of the abdominal part, (b) The 2D cross-sectional anatomic image, (c) Collaborative studying using a table-type virtual workbench

In this paper, we present a collaborative and immersive medical education system in a virtual workbench environment. A group of learners can be immersed into the study to understand the features of the subjects of interest more clearly in the wide semi-immersive virtual space. Moreover, learners can examine the dynamic features of the heart through the realistic motion simulation and also interactively observe the deformation of the soft tissues deformed by the efficient deformation approach in the proposed medical education system.

Acknowledgements

This work was partially supported by the Korean Ministry of Science and Technology under the NRL Program, and in part by the Korean Ministry of Information and Communication under the ITRC Program.

References

1. Weir, J., Abrahams, P.: 2nd Ed., Imaging Atlas of Human Anatomy. Mosby (1997)
2. Rost, R.J.: Clinical Anatomy for Medical Students. Lippincott Williams and Wilkins (2000)
3. Tiede, U., Schiemann, T., K.H.Hoehne: Visualizing the visible human. IEEE Computer Graphics and Applications **16** (1996) 7–9
4. Moody, D., Lozanoff, S.: Surfdriver: A practical computer program for generating three-dimensional models of anatomical structures. In: the 14th Annual Meeting of the American Association of Clinical Anatomists. (1997)
5. Teistler, M., Bott, O.J., Dormeier, J., Pretschner, D.P.: Virtual tomography: a new approach to efficient human-computer interaction for medical imaging. In: SPIE,Medical Imaging 2003: Visualization, Image-Guided Procedures, and Display. Volume 5029. (2003) 512–519
6. Santhanam, A.P., Fidopiastis, C.M., Hamza-lup, F., Rolland, J., Imielinska, C.: Physically-based deformation of high-resolution 3d models for augmented reality based medical visualization. In: Workshop AMI-ARCS 2004. (2004) 21–31
7. Sielhorst, T., Obst, T., Burgkart, R., Riener, R., Navab, N.: An augmented reality delivery simulator for medical training. In: Workshop AMI-ARCS 2004. (2004) 11–20
8. Tokuyasu, T., Oota, S., Asami, K., Kitamura, T., Sakaguchi, G., Koyama, T., Komeda, M.: Development of a training system for cardiac muscle palpation. In: Fifth International Conference on Medical Image Computing and Computer Assisted Intervention. (2002) 248–255
9. Zorin, D., Schröder, P., Sweldens, W.: Interpolating subdivision for meshes with arbitrary topology. In: ACM SIGGRAPH. (1996) 189–192

Extraction of Risk Factors by Multi-agent Voting Model Using Automatically Defined Groups

Akira Hara, Takumi Ichimura,
Tetsuyuki Takahama, and Yoshinori Isomichi

Faculty of Information Sciences, Hiroshima City University,
3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima, 731-3194, Japan
{ahara,ichimura,takahama,isomichi}@its.hiroshima-cu.ac.jp
http://www.chi.its.hiroshima-cu.ac.jp/

Abstract. In medical treatment, it is difficult to diagnose diseases directly from raw real value data obtained by medical examination for patients. For support of the task, it is necessary to transform the raw data into meaningful knowledge representations, which represent whether the characteristic symptoms of the disease are observed. In this research, we aim to acquire such multiple risk factors automatically from the medical database. We consider that a multi-agent approach is effective for extracting multiple factors. In order to realize the approach, we propose a new method using an improved Genetic Programming method, Automatically Defined Groups (ADG). By using this method, multiple risk factors are extracted, and the diagnosis is performed through multi-agent cooperative voting. We applied this method to the coronary heart disease database, and showed the effectiveness of this method.

1 Introduction

Recently, patient diagnostic data in hospitals have been accumulated in a database through the advance of information technology. We have been studying about the extraction of diagnostic rules from database by using an improved genetic programming. In our previous research[1], we succeeded in extracting multiple diagnostic rules represented by logical expressions. Both general rules and exceptional rules could be extracted by our method. However, we need to improve the recognition rate.

On the other hand, Oeda *et al.*[2] showed a good recognition rate for the same database by using a immune multi-agent neural network. However, it is difficult to extract knowledge from the acquired networks.

It is important to realize both improvement of the prediction accuracy and acquisition of the useful and comprehensible knowledge. We are now planning to develop a system with the advantages of the both methods. In the system, first, local characteristic symptoms of disease are extracted from the manifold viewpoints. In this first stage, comprehensible and useful knowledges about risk

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1218–1224, 2005.

factors are acquired by using an improved genetic programming. In the next stage, the final diagnosis is performed by the synthesis of the acquired multiple factors. Neural networks will be useful for this process. In actual medical treatment, doctors check the multiple manysided symptoms which may result from the disease. They make a final diagnosis based on the relation of observed multiple symptoms.

In this paper, we focus on the first stage, and propose the method for acquisition of multiple risk factors. We apply the method to the coronary heart disease database, and examine the effectiveness of this method.

There are some problems for realizing the system. We do not know how many factors are needed for classification of disease beforehand. Moreover, we do not know which test item is effective in classification and what is the appropriate threshold value of the test item for meaningful classification. In order to solve the problems, we use an improved Genetic Programming method, Automatically Defined Groups (ADG)[1]. Next, we describe this method.

2 Automatically Defined Groups

We use a multi-agent approach to extract multiple risk factors. That is, each agent discovers a risk factor, and a provisional diagnostic result is performed by multi-agent cooperation.

ADG is a method to optimize both the grouping of agents and the program of each group in the process of evolution. By grouping multiple agents, we can prevent the increase of search space and perform an efficient optimization. Moreover, the acquired group structure is utilized for understanding how many roles are needed and which agents have the same role. That is, the following three points are automatically acquired by using ADG.

- How many groups (roles) are required to solve the problem?
- Which group does each agent belong to?
- What is the program of each group?

A team that consists of all agents is regarded as one GP individual. One GP individual maintains multiple trees, each of which functions as a specialized program for a distinct group. We define a group as the set of agents referring to the same tree for the determination of their actions. All agents belonging to the same group use the same program.

Generating an initial population, agents in each GP individual are divided into random groups. Basically, crossover operations are restricted to corresponding tree pairs. However, we consider the sets of agents that refer to the trees used for the crossover. The group structure is optimized by dividing or unifying the groups according to the relation of the sets. Individuals search solutions as their group structures gradually approach the optimal structure.

The concrete processes are as follows: We arbitrarily choose an agent to two parental individuals. A tree referred to by the agent in each individual is used for crossover. We use T and T' as expressions of these trees, respectively. In

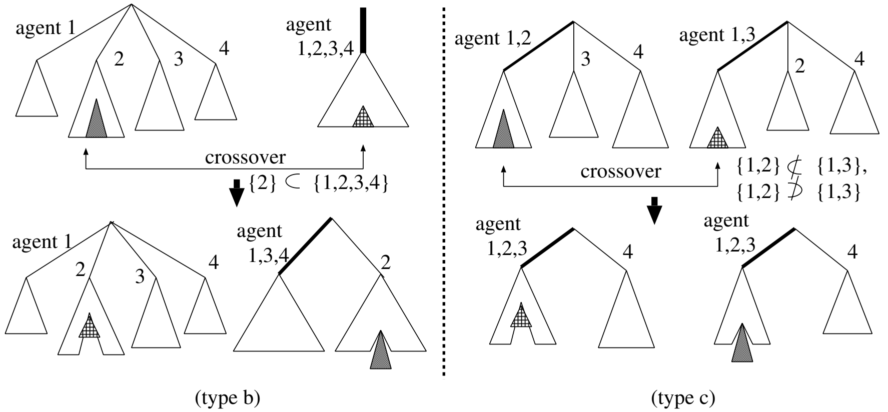


Fig. 1. Examples of crossover

each parental individual, we decide a set $A(T)$, the set of agents that refer to the selected tree T. When we perform a crossover operation on trees T and T', there are the following three cases.

(type a) If the relation of the sets is $A(T) = A(T')$, the structure of each individual is unchanged.

(type b) If the relation of the sets is $A(T) \supset A(T')$, the division of groups takes place in the individual with T, so that the only tree referred to by the agents in $A(T) \cap A(T')$ can be used for crossover. The individual which maintains T' is unchanged. Fig. 1 shows an example of this crossover.

(type c) If the relation of the sets is $A(T) \not\supset A(T')$ and $A(T) \not\subset A(T')$, the unification of groups takes place in both individuals so that the agents in $A(T) \cup A(T')$ can refer to an identical tree. Fig. 1 shows an example of this crossover.

We expect that, by using this method, the search works efficiently and the adequate group structure is acquired.

3 Extracting Risk Factors Using ADG

3.1 Coronary Heart Disease Database

In this section, we perform knowledge acquisition from real medical data. We use the database on coronary heart diseases[3]. Data in the coronary heart disease database are divided into two classes: non-coronary heart disease cases (non-CHD) and coronary heart disease cases (CHD). Each patient's disorder is diagnosed according to the results of eight test items. The eight items tested are Cholesterol (TC), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Left Ventricular Hypertrophy (LVH), Origin (ORIGIN), Education (EDUCATE), Smoking (TOBACCO), and Drinking (ALCOHOL).

The original results of some test items are provided as the real values with various ranges. So, we normalize each value. We find the maximum value (max) and minimum value (min) on each item in training data set, and i-th item's value x_i is normalized to X_i as follows:

$$X_i = (x_i - min_i)/(max_i - min_i)$$

We consider that categorical data such as EDUCATE and TOBACCO are also roughly in order of grade or frequency. Therefore we treate with the categorical data in the same manner as the numerical data.

In this research, we intend to construct a diagnostic system which can classify data into the appropriate class based on these eight tests.

3.2 How to Apply ADG to Extraction of Risk Factors

Even if the name of disease is the same, the individual variation may be observed in appearance of symptoms, or the symptoms may change by the growth of the disease. Therefore, the classification rule is not necessarily represented by a single rule. It is necessary to check symptoms from various angles and to diagnose the disease by the synthesis of them.

In order to judge whether each data is regarded as CHD case, we utilize a decision making by a multi-agent voting model. Each agent takes charge of the detection of a local symptom that should be considered for the classification of the disease. Each agent's tree structural logical expression represents the risk factor (the local symptom). Each agent examines whether its logical expression is true for the patient's data. If its logical expression returns true, the agent gives its vote to the approval for classifying the patients as the disease. If the number of approval is over a predefined threshold, the data is regarded as the disease.

The details of the process are as follows: Multiple trees in an individual of ADG represent the respective logical expressions. The logical expression is made by the conjunction of multiple terms. Each term is the combination of a test item and the value which can be taken. First, the numerical threshold values are generated with a uniform distribution between 0.0 and 1.0. The values are adjusted by mutation operations. The following expression is an example.

$$\text{Factor for CHD} : (TC > 0.51) \wedge (TC < 0.68) \wedge (DBP > 0.49)$$

Each data in the training set is input to all trees in the individual. Then, calculations are performed to determine whether the data satisfy each logical expression. Each tree returns true or false for the input data. Each agent receives the output of the tree which it refers to. The results is used for voting. Every agent votes based on its own tree's output. Fig.2 shows the concept of this model. This figure represents an individual which consists of five agents. The data in this figure is regarded as the disease, because approval votes are in majority. From another point of view, the number of agents in each tree represents the weight of the risk factor.

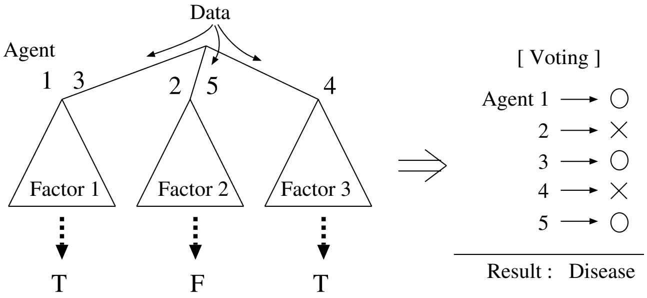


Fig. 2. Diagnosis by multi-agent voting model

ADG individuals are optimized so that the number of correct diagnosis increases. In addition, for the minute evaluation of individuals, we consider the number of approvals for misrecognition data. Even if a CHD data is missed, the number of approvals should be more. In contrast, if a non-CHD data is recognized as CHD by mistake, the number of approvals should be less.

To satisfy the requirements mentioned above, fitness f is calculated by the following equation. We maximize f by evolution.

$$f = \frac{true\_positive}{N_{CHD}} + \alpha \frac{true\_negative}{N_{nonCHD}} + \beta miss\_target + \gamma\; misrecognition \quad (1)$$

where, N_{CHD} and N_{nonCHD} represent the number of CHD cases and non-CHD cases in database respectively. *true_positive* is the number of successful data in the target CHD data. *true_negative* is the number of successful data in the non-CHD data. Moreover, *miss_target* is an average approval rate for the CHD data that have been missed. *misrecognition* is an average disapproval rate for the non-CHD data that have been regarded as CHD by mistake.

By evolution, More agents, the number of which is over the judgment threshold, learn to return true for CHD cases, and few agents return true for non-CHD cases. Therefore, the factor with more agents is the typical and crucial risk factor.

The following points are regarded as the advantages of ADG.

- ADG enables us to get multiple factors of disease and to make a final diagnosis by using the facotrs.
- It is easy to judge by the number of agents whether the acquired factors are crucial ones.
- It is easy to understand the acquired factor, because the factor is expressed by the logical expression and the threshold values of the test items are also acquired.

Table 1 shows GP functional and terminal symbols. We impose constraints on the combination of these symbols. Terminal symbols do not enter directly in the arguments of the `and` function. Test items such as `TC` enter only in arg0 of `gt` and `lt`. Real values enter only in arg1. Crossover and mutation that break the constraints are not performed.

Table 1. GP Functions and Terminals

Symbol	#args	functions
`and`	2	arg0 ∧ arg1

`gt`	2	if (arg0 > arg1)
		return T else return F

`lt`	2	if (arg0 < arg1)
		return T else return F

`TC`, `SBP`, ...	0	normalized test value
`0.0` – `1.0`	0	real value

4 Results

In this section, we describe the detail of an experiment using Train_A, which are consisted of 6500 CHD cases, and 6500 non-CHD cases. The parameter settings of ADG are as follows: Population size is 500, crossover rate is 0.9, mutation rate per individual is 0.95, group mutation rate is 0.02, and the number of agents is 50. The threshold value (the number of approvals) for the classification by the voting is 30% of the number of all agents. The respective weights in equation(1) are $\alpha = 1.0$, $\beta = 0.05$, and $\gamma = 0.01$.

As a result, 50 agents in the best individual are divided into 21 groups. We show some of acquired factors that correspond to the tree structural programs in the best individual. Factors are arranged according to the number of agents referring the each factor, and each terminal real value is transformed to original range.

Factor 1 (6 Agents): (LVH = 1)
Factor 2 (4 Agents): (SBP > 148)
Factor 3 (4 Agents): (TC > 282) ∧ (DBP > 76) ∧ (LVH = 0) ∧ (TOBACCO > 2)
Factor 4 (4 Agents): (SBP > 188) ∧ (DBP < 99)
Factor 5 (3 Agents): (SBP > 126)
Factor 6 (3 Agents): (DBP > 70)
Factor 7 (3 Agents): (DBP > 110) ∧ (DBP < 147) ∧ (LVH = 0)
Factor 8 (2 Agents): (SBP > 118)
Factor 9 (2 Agents): (SBP > 160) ∧ (DBP > 97)
Factor 10 (2 Agents): (TC > 206) ∧ (DBP > 82)
Factor 11 (2 Agents): (TC > 256) ∧ (TC < 358)
Factor 12 (2 Agents): (EDUCATE < 2)

This results shows that any risk factors do not have the number of agents that exceeds the diagnostic threshold solely. However, the number of agents of factor 1 is closer to the threshold than the other factors. If other slight factors are also valid, it can be concluded that the patient has the disease.

Besides, the factor 2 has the same test item as the factor 5 or the factor 8. However, The reference value of the item in factor 8 is more moderate than that in factor 2. Therefore the influence of the factor 8 on the final diagnosis is light, and the check of symptoms from other angles are also needed. On the contrary, if the factor 2 becomes true, the factor 5 or 8 also becomes true. In the case, nine agents at least give their votes to the approval. Thus, we can set two or more reference values to a test item, and utilize them for the diagnoses.

By the way, we have performed several trials of this method. The sets of multiple factors acquired by respective trials are not necessarily identical. This result comes from a diversity of cooperative behavior. We will have to examine the stability of solutions.

The classification accuracies in the above experiment are 70.0% for 13000 training cases and 69.3% for 13000 test cases. As preliminary experiments for the final classification by the synthesis using neural networks, we use two sets of risk factors acquired by different trials as inputs for a feed-forward neural network. By a back-propagation learning, the classification accuracy becomes 79.7% for the training cases. This result suggests that multi-agent voting model is effective for the extraction of comprehensible factors and a provisional classification, and neural networks are promising for a final accurate classification based on the acquired factors.

5 Conclusions and Future Work

In this research, we proposed a new method using ADG for the purpose of the extraction of multiple risk factors. In this method, the weights of respective factors also can be acquired by multi-agent cooperation through the voting model. We showed the effectiveness of this method by the application to medical data.

We have to investigate the usefulness of extracted factors from the viewpoint of health care. In addition, we need to improve the classification accuracy. We will study how to synthesize risk factors by using neural networks, which seems to be more flexible than the voting model.

Acknowledgments

This work was supported by a Grant-in-Aid for Young Scientists (B) (No. 15700199) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. In addition, this was also supported by a Hiroshima City University Grant for Special Academic Research (General Study).

References

1. A. Hara, T. Ichimura, T. Takahama and Y. Isomichi: “Extraction of Rules from Coronary Heart Disease Database Using Automatically Defined Groups”, *Proc. The Eighth Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES'2004)*, Vol.2, pp.1089-1096 (2004)
2. S. Oeda, T. Ichimura and K. Yoshida: “Immune Multi Agent Neural Network and Its Application to the Coronary Heart Disease Database”, *Proc. The Eighth Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES'2004)*, Vol.2, pp.1097-1105 (2004)
3. M. Suka, T. Ichimura and K. Yoshida: “Development of Coronary Heart Disease Database”, *Proc. The Eighth Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES'2004)*, Vol.2, pp.1081-1088 (2004)

Representing Association Classification Rules Mined from Health Data*

Jie Chen[1], Hongxing He[1], Jiuyong Li[4], Huidong Jin[1], Damien McAullay[1], Graham Williams[1,2], Ross Sparks[1], and Chris Kelman[3]

[1] CSIRO Mathematical and Information Sciences
GPO Box 664, Canberra ACT 2601, Australia
{Jie.Chen,Hongxing.Heinst,Huidong.Jin}@csiro.au
{Damien.McAullay,Graham.Williams,Ross.Sparks}@csiro.au
[2] Current address: Australian Taxation Office, Australia
Graham.Williams@togaware.com
[3] National Centre for Epidemiology and Population Health,
The Australian National University, Australia
Chris.Kelman@anu.edu.au
[4] Department of Mathematics and Computing,
University of South Queensland, Australia
jiuyong@usq.edu.au

Abstract. An association classification algorithm has been developed to explore adverse drug reactions in a large medical transaction dataset with unbalanced classes. Rules discovered can be used to alert medical practitioners when prescribing drugs, to certain categories of patients, to potential adverse effects. We assess the rules using survival charts and propose two kinds of probability trees to present them. Both of them represent the risk of given adverse drug reaction for certain categories of patients in terms of risk ratios, which are familiar to medical practitioners. The first approach shows risk ratios when all rule conditions apply. The second presents the risk associated with a single risk factor with other parts of the rule identifying the cohort of the patient subpopulation. Thus, the probability trees can present clearly the risk of specific adverse drug reactions to prescribers.

1 Introduction

Data mining usually involves extracting actionable knowledge from databases. Thus, understanding and evaluating the discovered patterns become increasingly important, especially in health applications. Systematic monitoring of adverse drug reactions is important for both financial and social reasons. At present, the early detection of unexpected adverse drug reactions relies on a national

* The authors acknowledge the Commonwealth Department of Health and Ageing, and the Queensland Department of Health for providing data and support for this research.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1225–1231, 2005.

spontaneous reporting system and collated statistics from overseas agencies [1, 2]. However, the recent availability of a population-based prescribing dataset, such as the Pharmaceutical Benefits Scheme (PBS) data in Australia, when linked to hospital admissions data, provides a unique opportunity to detect rare adverse drug reactions at a much earlier stage before many patients are affected. This paper focuses on identifying the factors, which increase the risk of the adverse drug reaction, directly from large linked health data rather than spontaneous reporting databases.

Prescribed drugs are recorded in PBS data based on the Anatomical and Therapeutic Classification (ATC) system. Adverse events are recorded in hospital data using ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) code. Three case studies have been identified by experts from the Therapeutic Goods Administration, Australia. ACE inhibitors[1] usage associated with Angioedema will serve as the main case study to illustrate our method in this paper. In our data, the distribution of classes with and without adverse events is highly unbalanced due to the intrinsic nature of adverse drug reactions. Moreover, rules identified may be used to alert medical practitioner in their prescription of drugs to certain categories of patients, who are vulnerable to some adverse drug effects. It is therefore essential to present the knowledge to medical practitioners in a form easy to understand and interpret. To address this health data mining problem, we first modify the Optimal Class Association Rule Mining Algorithm [4] to discover rules which identify patient subgroups with a high proportion of patients with target events. We further propose two kinds of tree representation for mined rules to help them and potential users to gain understanding of the rules.

2 Association Classification for Unbalanced Classes

Traditional association classification approaches search for the rules represented by patterns which have high global support and high confidence. Since the "normal" group comprises more than 99% of all cases in the dataset, the class of interest(Class 1 defined in Section 3) is given little attention by these approaches. In this paper, we modify the Optimal Class Association Rule Mining Algorithm [4] by introducing local support and risk ratio to identify higher risk patient groups of adverse drug reaction events. The support in minor class is called *local support* defined as $\frac{sup(A \rightarrow c)}{sup(c)}$. Here $sup(c)$ and $sup(A \rightarrow c)$ represent the support (or proportion) of Class c in the whole population and the support of pattern A in Class c respectively. Minimum local support can be used as a parameter of the algorithm to specify the minimum fraction of population of interest in each class of the unbalanced dataset. We propose to use the *Risk Ratio* as measure of interestingness for pattern mining, which is represented by $RR(A \rightarrow c) = \frac{lsup(A \rightarrow c) sup(\overline{A})}{lsup(\overline{A} \rightarrow c) sup(A)}$.

[1] Angioedema is a swelling that occurs beneath the skin rather than on the surface [3]. There are a number of case series in the literature demonstrating that ACE inhibitor-related angioedema is responsible for as many as 40% of angioedema episodes [3]

Table 1. List of variables used for association classification

Variable	Values	Variable	Values
Gender	m,f	Alimentary tract metabolism	0,1
Age group	1,2,3,4	Blood and blood forming organs	0,1
Indigenous	0,1	Cardiovascular systems	0,1
Sickness(bed days)	1,2,3	Dermatologicals	0,1
Hosp. Neoplasm Flag	0,1	Genito urinary system and sex hormones	0,1
Hosp. Diabetes Flag	0,1	Systematic hormonal preparations	0,1
Hosp. Mental Health Flag	0,1	General anti-infective for systematic use	0,1
Hosp. Circulatory Flag	0,1	Antineoplastic and immunimodulating agents	0,1
Hosp. Ischaemic Heart Disease Flag	0,1	Musculo-skeletal system	0,1
Hosp. Respiratory Flag	0,1	Nervous system	0,1
Hosp. Asthma Flag	0,1	Antiparasitic products insecticides and repellents	0,1
Hosp. Musculoskeletal Flag	0,1	Respiratory system	0,1
Total Scripts	0,1,2	Sensory organs	0,1
Class	0,1	Various	0,1

The risk ratio defines the relative risk (belonging to Class 1) of the patients identified by rule A with respect to the majority of patients [5, p. 35]. $\overline{A}$ denotes the absence of pattern A.

3 Data Preparation and Feature Selection

The Queensland Linked Data Set [6] links hospital admissions data from Queensland Health with the pharmaceutical prescription data from the Commonwealth Department of Health and Ageing, providing a de-identified dataset for analysis. For the implementation of the mining task, we chose to extract profile data for all patients exposed to the drug of interest in a 180 day window, which was selected using domain knowledge. The patients are further partitioned into two classes (Class 1 and 0). The patients in Class 1 are such patients that have taken the target drugs (e.g. ACE inhibitors) within the time window prior to the first adverse drug reaction event, and other patients are in Class 0. Features selected for the profile of each patient are described below.

From the hospital data, demographic variables such as age, gender, indigenous status, postcode, the total number of bed days and the eight hospital diagnosis flags are extracted. The hospital diagnosis and the total number of bed days can be used to infer the health status of an individual. From the PBS data, another 15 variables (including such variables as the total number of scripts of the specified drug and the 14 ATC level-1 drugs) were extracted. The "Total scripts" is used to indicate how long an individual has been exposed to the drug (because each script usually provides medication for one month). The 14 ATC level-1 drug categories may be useful in measuring adverse drug reactions caused by interactions between the specified drug and other drugs.

Table 1 lists the variables representing the profiles of patients. We chose some variables in the profiles in applying the association classification algorithm. "Age", "Bed days" and "Total scripts" are discretised because the algorithm requires all the variables take only a set of discrete values. Since the aim of the algorithm is to identify the group of patients who are more likely to have an adverse drug reaction than the general population, we choose these variables, which are most commonly considered as important for their health and wellbeing. We consulted medical practitioners to incorporate their knowledge in our study.

There are limitations in selecting best variables as our dataset is not from survey data, e.g. some desirable variables such as life style information can not be obtained.

4 Representing Association Classification Rules

Usually when the modified optimal class association rule mining algorithm is applied to identify the high risk groups, a large number of rules with risk ratio greater than 2.0 are generated. The exceptional rules (risk ratio is less than 1.0) could be interesting in identifying lower than general risk groups. However, they are not primary objectives of the current study and therefore ignored. We could not present hundreds of rules to medical experts for inspection. Furthermore, most of them are correlated and provide similar information. We can select rules by an effective method. Let all generated rules match all records in the dataset and only keep the rule with the highest risk ratio for each record. This will reduce the number of rules significantly.

The five rules with highest risk ratio for the ACE inhibitors and angioedema case study are listed below:

Rule 1: RR = 3.9948
- Gender = Female
- Hospital Circulatory Flag = Yes
- Usage of Drugs in category "Various" = Yes

Rule 2: RR = 3.8189
- Age > 60
- Usage of drugs in category of "Genito urinary system and sex hormones" = Yes
- Usage of drugs in category of "Systematic hormonal preparations" = Yes

Rule 3: RR = 3.4122
- Usage of drugs in category of "Genito urinary system and sex hormones" = Yes
- Usage of drugs in category of "General anti-infective for systematic use" = Yes
- Usage of drugs in category of "Nervous system" = No

Rule 4: RR = 3.3269
- Gender = Female
- Age group in [40, 59]
- Total bed days $\geq$ 15

Rule 5: RR = 3.2605
- Usage of drugs in category of "Alimentary tract metabolism" = No
- Usage of drugs in category of "Genito urinary system and sex hormones" = Yes
- Usage of drugs in category of "General anti-infectives for systematic use" = Yes

For each rule discovered, we conduct further evaluation, e.g., the survival analysis and its significance test [5, pp. 159-169]. In addition, we use the log-rank test, a formal measure of the strength of evidence that two populations have different lifetimes. Fig. 1 presents the estimated survival functions of the subgroup described by Rule 5 (the one within the filled region) and the other patients (within the shaded region). The filled region and the shaded region indicate their confidence intervals, respectively. Clearly, for the age range from 60 to about 80, the subgroup indicated by Rule 5 has significantly higher probability of hospital admission for angioedema than the other patients. The P-value of the log-rank test is 5.0583e-09, which suggests that the sub-group described by Rule 5 is overwhelmingly different from the other patients. Similar interesting results are also found in other rules [6].

The rules identified by the association classification algorithm provide useful knowledge to medical practitioners, and can serve as a reference in their prescription of drugs to the patients. The patients' characteristics can be compared

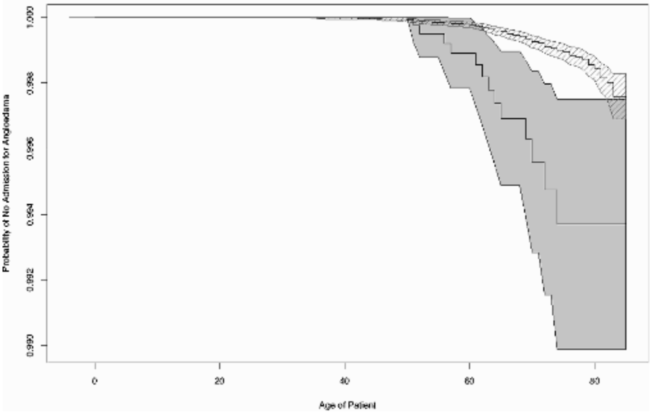


Fig. 1. Fleming-Harrington survival analysis of Rule 5

to the rules to evaluate their risk to the suspected adverse drug reaction. However, the rules presented above may not provide enough information for clinical use. The further breakdown of the risks caused by individual factors provides important information in their assessment of the risk. Therefore we employ a tree structure to visualise the rules mined. A variable value pair is presented at each node of the tree. The information on the support of the population, its percentage and risk ratio is presented on each node. The branch to the right of the node lists the information for complementary population. The level down of each node gives another split of population using a new variable value pair. As an example, Rule 1 is presented as a tree in Fig. 2. Note that most commonly used multiple logistic regression models can be used and similar tree structures could be obtained accordingly. However, the presentation method can rank rules according to their relative risks automatically to avoid time consuming model analysis work.

According to Fig. 2, female users of ACE inhibitors are 1.54 times more likely to have angioedema than the population average. For those female patients who have a circulatory disease, the likelihood increases to 1.82. For those who are female, have a circulatory disease, and also have taken drugs falling in the "Various" category (the 14th ATC level-1 drug category), the likelihood increases further to 4.0. The tree presentation highlights how the risk ratio changes with each individual component. Further stratifications may help to make rules more adaptable in clinical decisions. Alternatively, we can define the risk ratio at each node to be relative to the population of its parent node. Accordingly the risk ratio at each node is expressed by $RR(A \to C \mid U) = \frac{lsup(A \bigcap U \to C) sup(\overline{A} \bigcap U)}{lsup(\overline{A} \bigcap U \to C) sup(A \bigcap U)}$, where U is the rule on the parent node.

The tree presentation of the same rule using the alternative definition of risk ratio is presented in Fig. 3. According to Fig. 3, female users of ACE inhibitors are 1.54 times more likely to have angioedema than the population average. For female patients, the patients who have a circulatory disease, are 1.92 times more

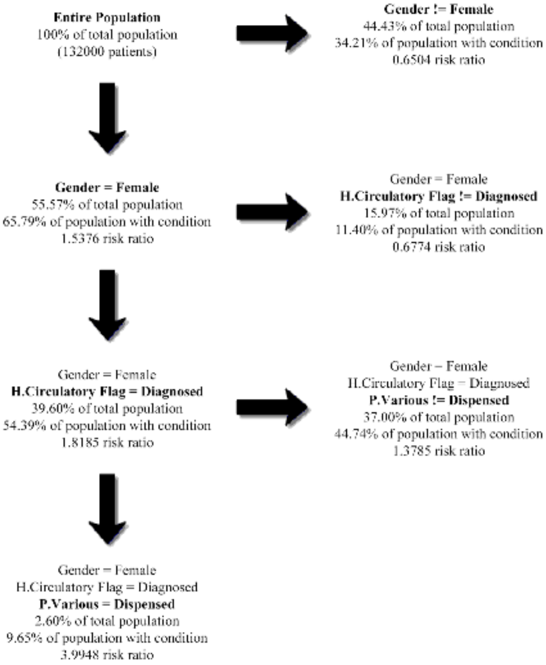

Entire Population
100% of total population
(132000 patients)
Gender != Female
44.43% of total population
34.21% of population with condition
0.6504 risk ratio
Gender = Female
55.57% of total population
65.79% of population with condition
1.5376 risk ratio
Gender = Female
H.Circulatory Flag != Diagnosed
15.97% of total population
11.40% of population with condition
0.6774 risk ratio
Gender = Female
H.Circulatory Flag = Diagnosed
39.60% of total population
54.39% of population with condition
1.8185 risk ratio
Gender – Female
H.Circulatory Flag = Diagnosed
P.Various != Dispensed
37.00% of total population
44.74% of population with condition
1.3785 risk ratio
Gender = Female
H.Circulatory Flag = Diagnosed
P.Various = Dispensed
2.60% of total population
9.65% of population with condition
3.9948 risk ratio

Fig. 2. The first tree presentation of Rule 1

likely to develop angioedema than other female patients. The female patients with a circulatory disease, and who have used drugs in the “Various” category are 3.06 times more likely than female patients with a circulatory disease but not taking drugs in that category. However, we need to keep in mind that the rule presentation can help doctors to be alert in prescribing medicine to patients with certain characteristics. The indication becomes more complex when patients have multiple diseases such as asthma and diabetes etc.

5 Conclusion

In this paper, we have applied a modified association classification algorithm to health data to explore risk factors associated with adverse drug reactions. We assessed the discovered rules using survival charts and introduced two tree-type presentations to present risk factors in a comprehensible way. The tree presentations are able to demonstrate the heightened risks due to a combination of risk factors as well as due to a single risk factor. Thus, they provide an effective way for medical practitioners to interpret clearly the risk factors for prescribing certain drugs to specific patient sub-groups. The consequence of this should be more effective use of medicines and reduced morbidities or costs from adverse drug events. Such knowledge could be readily implemented in electronic prescribing systems.

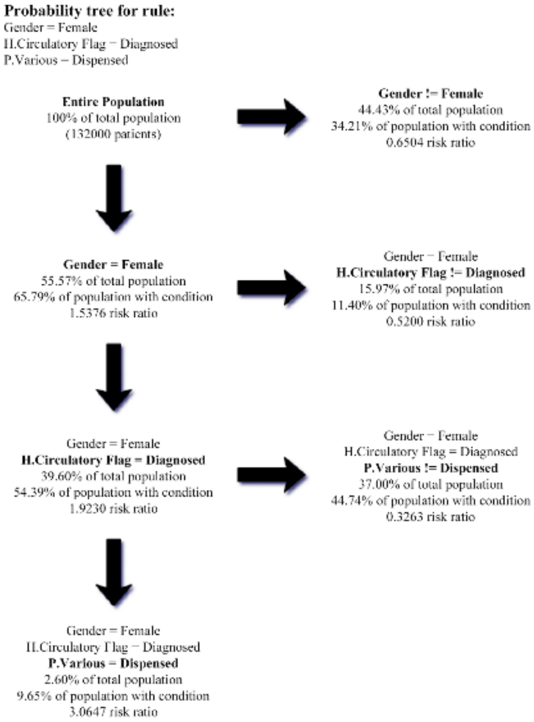

Probability tree for rule:
Gender = Female
H.Circulatory Flag = Diagnosed
P.Various = Dispensed
Entire Population
100% of total population
(132000 patients)
Gender != Female
44.43% of total population
34.21% of population with condition
0.6504 risk ratio
Gender = Female
55.57% of total population
65.79% of population with condition
1.5376 risk ratio
Gender = Female
H.Circulatory Flag != Diagnosed
15.97% of total population
11.40% of population with condition
0.5200 risk ratio
Gender = Female
H.Circulatory Flag = Diagnosed
39.60% of total population
54.39% of population with condition
1.9230 risk ratio
Gender = Female
H.Circulatory Flag = Diagnosed
P.Various != Dispensed
37.00% of total population
44.74% of population with condition
0.3263 risk ratio
Gender = Female
H.Circulatory Flag = Diagnosed
P.Various = Dispensed
2.60% of total population
9.65% of population with condition
3.0647 risk ratio

Fig. 3. The second tree presentation of Rule 1

References

1. David M. Fram, June S. Almenoff, and William DuMouchel. Empirical bayesian data mining for discovering patterns in post-marketing drug safety. In *Proceedings of KDD 2003*, pages 359–368, 2003.
2. Harvey J. Murff, Vimla L. Patel, George Hripcsak, and David W. Bates. Detecting adverse events for patient safety research: a review of current methodologies. *Journal of Biomedical Informatics*, 36(1/2):131–143, 2003.
3. M. Reid, B. Euerle, and M. Bollinger. Angioedema, 2002. http://www.emedicine.com/med/topic135.htm.
4. J. Li, H. Shen, and R. Topor. Mining the optimal class association rule set. *Knowledge-Based Systems*, 15(7):399–405, 2002.
5. Stephen C. Newman. *Biostatistical Methods in Epidemiology.* John Wiley & Sons, July 2001.
6. Graham Williams, Hongxing He, Jie Chen, Huidong Jin, Damien McAullay, Ross Sparks, Jisheng Cui, Simon Hawkins, and Chris Kelman. QLDS: Adverse drug reaction detection towards automation. Technical Report CMIS 04/91, CSIRO Mathematical and Information Sciences, Canberra, 2004.

Leximancer Concept Mapping of Patient Case Studies

Marcus Watson[1], Andrew Smith[1], and Scott Watter[2]

[1] Key Centre for Human Factors and Applied Cognitive Psychology,
The University of Queensland, Australia
{mwatson,asmith}@humanfactors.uq.edu.au
http://www.humanfactors.uq.edu.au/
[2] Department of Psychology, McMaster University, Canada
watter@mcmaster.ca

Abstract. Quantitative databases are limited to information identified as important by their creators, while databases containing natural language are limited by our ability to analyze large unstructured bodies of text. Leximancer is a tool that uses semantic mapping to develop concept maps from natural language. We have applied Leximancer to educational based pathology case notes to demonstrate how real patient records or databases of case studies could be analyzed to identify unique relationships. We then discuss how such analysis could be used to conduct quantitative analysis from databases such as the Coronary Heart Disease Database.

1 Introduction

Databases only contain information that the creators of the database consider potentially relevant to the topic. For example the information contained within the Coronary Heart Disease Database reflects physiological variables and population information that has historically correlated with heart disease [1]. These databases are limited by the richness of the information contained; for example they are unlikely to provide information on how a drug for arthritis may increase the occurrence of cardiac arrest. The information required to identify such relationships is likely to be found in patient records and case studies; however, inconsistencies in how patient records and case studies are compiled means that such material is more ambiguous than the data found in the Coronary Heart Disease Database. The sheer quantity of patient records provides a significant obstacle to overcome when searching for good quality evidence on which to base health strategies [2]. Although some progress has been made to include narrative descriptions from case studies and patient records, the rate that the patient information is generated far outstrips our ability to classify the information into databases.

Databases such as the Australian Incident Monitoring Study (AIMS) go beyond physiological variables and population information to include adverse case descriptions [3]. The use of written descriptions of patient events in these databases combined with physiological information has the potential to identify complex relationships that are unlikely to be detected through a database containing physiological variables and population information alone. However although the AIMS database and others like it are rich in information, they are limited in three ways. (1) They rely

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1232–1238, 2005.

on voluntary reporting of adverse incidents; therefore, the database may not reflect the rate of actual adverse incidents occurring if under-reporting occurs. (2) The number of cases may be small due to the low occurrence and low rate of reporting of adverse patient events. (3) Human analysts may be subject to influences which they are unable to report [4], which are likely to influence the interpretation of patient incident narrative.

Rapid analysis of databases containing large numbers of patient records and case studies may be possible if we are able to semantically map the text contained in the raw patient records or case studies. We will describe Leximancer, a tool that uses semantic mapping to develop concept maps from natural language and then provide an example of how we have applied Leximancer to some educational based pathology case notes. We will demonstrate Leximancer using the educational based case studies to show how real patient records or databases of case studies could be analyzed to identify unique relationships not available on databases containing physiological variables and population information alone. We will then discuss how such analysis could be used to conduct quantitative analysis from databases such as the Coronary Heart Disease Database.

2 Leximancer Concept Mapping Tool

The Leximancer system is a method for transforming lexical co-occurrence information from natural language into semantic patterns in an unsupervised manner. The technology behind Leximancer is based on Bayesian theory where fragmented pieces of evidence can be used to predict what is actually happening in a system. In natural language, a word can be defined by its context in usage, as words tend to correlate with other words over a certain range within the text stream [5]. Leximancer employs semantic then relational co-occurrence information extraction to identify relationships in the documents. The algorithms are statistical but they employ nonlinear dynamics and machine learning.

Leximancer assists the analyst to discover new information from text, find patterns across datasets, and may assist in separating signals from noise. Using the Leximancer generated context maps and statistical outputs, analysts may be better positioned to efficiently gain insight and comprehension of large numbers of case studies or even patient records through the process of: (1) conducting semantic information retrieval of the key themes of records; (2) viewing bodies of data in a graphical format, and (3) navigating through the records whilst mining the text for deeper contextual associations [6, 7, 8]. Domain expertise is not required to construct the maps, but will enhance the interpretation.

Leximancer is able to automatically extract the most important concepts in a set of documents such as patient records or case studies that are processed. Smith and Humphreys [8] have undertaken validation of the Leximancer methods. The algorithms used are not only able to analyze well-structured text, but also text where the author has used dot points or short comments as often found in patient records. The algorithms allow for the retrieval of episodic text records using semantic representations of cue words, even when the initial cue words are not present in the text records.

We have already demonstrated the ability of Leximancer to process and conceptually analyze very large bodies of text, with a view to drastically reducing the length of time it would take humans to read and draw key themes from such documents. We have so far processed database of documents exceeding 300MB. It also provides a means of quantifying and displaying information to explore interesting conceptual features [6]. The concept maps produced by Leximancer showing the relative co-occurrence of concepts not only help users to find the information they seek but also support the discovery of unexpected relationships that may be relevant to the user's investigation. This is evident through the visualization of the strengths between concepts and the similarities in which concepts occur. Visually emergent concept groups are referred to as themes. Examples of these are highlighted in Figure 1 by the hand drawn circles. Identification of themes by analysts is facilitated by employing the hierarchy of concept connectedness. Each theme is traditionally characterized by the most highly connected concept in the thematic region.

Leximancer's automatic mapping process is likely to reduce expertise bias when interpreting a set of documents such as patient records or case studies. This is achieved through both the visualization of concept relationships on the maps and through the ability to drill down into the text to identify the reasons for the co-occurrence of concepts. Leximancer consequently has the potential to support analysts in dealing with information sources such as the AIMS database without the need to rely on key word identification. More importantly, Leximancer could be applied to raw patient records which would avoid the problem of selective case reporting while still allowing the analyst to cope with a vast quantity of patient records. To demonstrate this we have used the online teaching case database from the Department of Pathology at University of Pittsburgh School of Medicine.

3 Applying Leximancer to Patient Case Notes

Leximancer is capable of dealing with vast quantities of patient records; however, the availability of patient records is understandably very limited. It is important that the database used is in the public domain so that other researchers can make comparisons. We therefore have chosen to use a database of pathology teaching material available on the www. The pathology case study database is not designed for research; however, it does provide enough cases to demonstrate how Leximancer could be applied to similar case study databases or even a database of patient records. At the time of our Leximancer analysis, the database contained 421 case studies that covered several categories reflecting the scope of clinical and diagnostic expertise at the Department of Pathology at the University of Pittsburgh School of Medicine <http://path.upmc.edu/index.html>. The initial phase of the Leximancer analysis involved developing an automatic concept map of the database. The automatic concept map was used to identify the central themes of case studies. Two concepts were then selected to demonstrate how Leximancer can be used to extract particular themes from the database. The first concept selected was Tumor, which was identified as a central concept within the database. The second concept was the Heart which, while present on the automatic concept map, is not a central concept of the case studies contained within the database.

The map of concepts relating to Tumor is shown in Figure 1. Each word on the map indicated a concept related to Tumor that the authors discuss in the case studies. Each concept has a colored circle (the small grey dots at the center of each word) indicating the occurrences of the concept relative to others in the case studies. The size of these circles indicates the centrality of the concepts in the text. The grey hand drawn circles indicate the groupings of concepts that can be aggregated to form themes. Identifying where themes occur and productively labeling them can be difficult without expertise in the domain being investigated. The labeling of the themes in Figure 1 is based upon the third author's medical expertise. The relationship between concepts within each theme can then be used to examine the case studies database. We will describe how this can be done in the second example, looking at the seeded concept of Heart. If this analysis was performed with real patient data, validation by multiple health care professionals would be required.

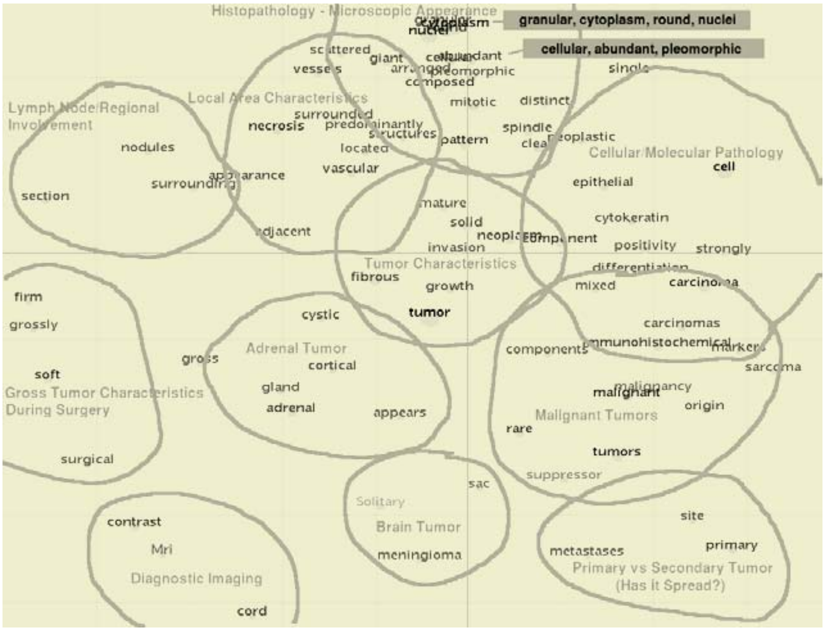


Fig. 1. Hand seeded Leximancer concept map for *Tumor* from the 421 cases studies

The use of hand seeded Leximancer concepts allows the analyst to search a large database of text for concepts and themes that may rarely appear in the body of work. For example if we were to examine all the patient records within a single hospital with an automatic Leximancer map; rarely occurring diseases are unlikely to appear in the top few hundred concepts. As maps containing more than a few hundred concepts become difficult to read it is important that the analyst can direct the type of information they desire to investigate. By hand seeding a concept, even one that rarely occurs, Leximancer can generate a map of the related concepts. The database we used here does not focus on issues related to coronary heart disease; however, we considered that there may be some relevant pathology information.

In the second analysis we examined how much information the database provides on coronary heart disease by seeding the concept Heart (Figure 2), in this instance by starting simply with the seed term 'heart'. Selection of the seed word was the only human intervention required – the rest of the taxonomy was discovered automatically. The same database reveals a very different map when compared to the map for Tumor. In this case, many of the theme groupings are more obvious as they form dis-

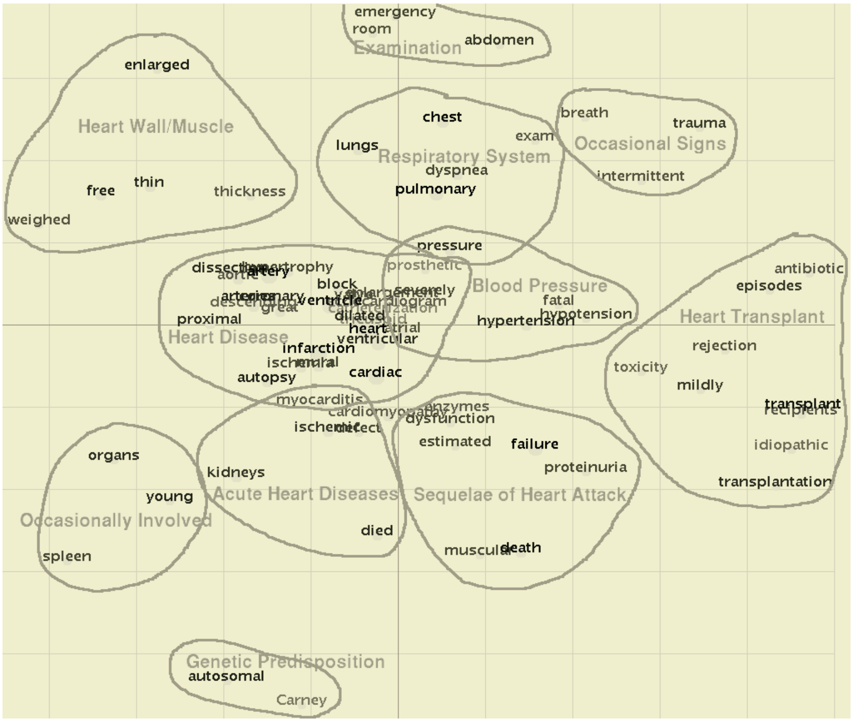
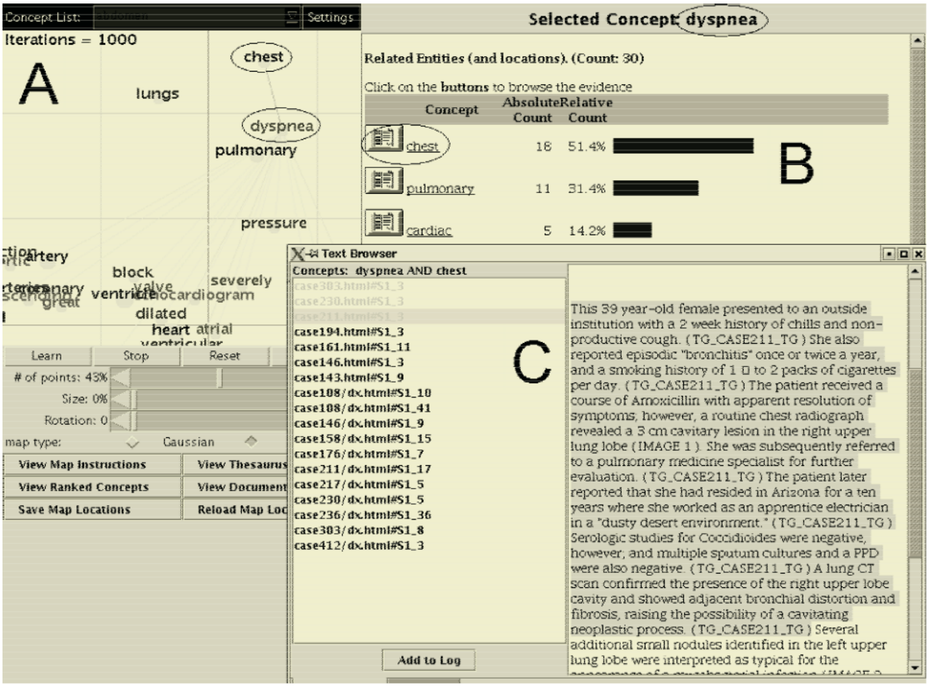


Fig. 2. Hand seeded Leximancer concept map for *Heart* from the 421 cases studies

tinct groups. If we were using a database containing a greater number of pathology cases, then we could effectively drill down into the case by examining the relationships in each theme. To give an example of how this can be done and why semantic relationships within case studies and patient records is so important we give one example of how the co-occurrence of concepts can be used.

We have selected two Respiratory System concepts “dyspnea” and ‘chest’ that illustrate how Leximancer can be used to identify case information that could be missed when large natural language databases are examined using key word searches. Figure 3 shows some of the features of Leximancer. (A) The left map shows a zoomed in view of the field of concepts for the Heart seeded map shown in Figure 2. (B) Leximancer extracts can be used to browse for occurrences or co-occurrences of specific concepts and relationships. For example, (B) shows the concepts that co-occur with the selected concept ‘dyspnea’. Selecting the button beside the related “chest” brings up the instances of where the concept co-occurs. (C) The text browser shows the retrieval of one of the 18 incidences of ‘dyspnea’ and ‘chest’ co-occurring in the 421 cases contained in the database.

In this particular incident of the co-occurring ‘dyspnea’ and ‘chest’, the word ‘chest’ does occur within the text; however, the word ‘dyspnea’ does not. It is evident that Leximancer has developed an accurate thesaurus for the word ‘dyspnea’ from the database by comparing this incident (the shaded area on the right of C in Figure 3) with the definition of dyspnea. Dyspnea is defined as abnormal or uncomfortable breathing in the context of what is normal for a person according to their level of



Selected Concept: dyspnea

Related Entities (and locations). (Count: 30)

Click on the **buttons** to browse the evidence

Concept	Absolute Count	Relative Count
chest	18	51.4%
pulmonary	11	31.4%
cardiac	5	14.2%

Concepts: dyspnea AND chest

case303.html#S1_3
case230.html#S1_3
case211.html#S1_3
case194.html#S1_3
case161.html#S1_11
case146.html#S1_3
case143.html#S1_9
case108/dx.html#S1_10
case108/dx.html#S1_41
case146/dx.html#S1_9
case158/dx.html#S1_15
case176/dx.html#S1_7
case211/dx.html#S1_17
case217/dx.html#S1_5
case230/dx.html#S1_5
case236/dx.html#S1_36
case303/dx.html#S1_8
case412/dx.html#S1_3

This 39 year-old female presented to an outside institution with a 2 week history of chills and non-productive cough. (TG_CASE211_TG) She also reported episodic "bronchitis" once or twice a year, and a smoking history of 1 □ to 2 packs of cigarettes per day. (TG_CASE211_TG) The patient received a course of Amoxicillin with apparent resolution of symptoms, however, a routine chest radiograph revealed a 3 cm cavitary lesion in the right upper lung lobe (IMAGE 1). She was subsequently referred to a pulmonary medicine specialist for further evaluation. (TG_CASE211_TG) The patient later reported that she had resided in Arizona for a ten years where she worked as an apprentice electrician in a "dusty desert environment." (TG_CASE211_TG) Serologic studies for Coccidioides were negative, however, and multiple sputum cultures and a PPD were also negative. (TG_CASE211_TG) A lung CT scan confirmed the presence of the right upper lobe cavity and showed adjacent bronchial distortion and fibrosis, raising the possibility of a cavitating neoplastic process. (TG_CASE211_TG) Several additional small nodules identified in the left upper lung lobe were interpreted as typical for the

Fig. 3. The Leximancer GUI showing the concept map, concept co-occurrence and text 'drill down' showing the co-occurrence of dyspnea and chest

fitness and exertion threshold for breathlessness [10]. A key word search of the database for dyspnea would not have identified that this incident is a relevant case.

Leximancer concepts based on patient records and/or case studies can be assessed in similar ways to quantitative data such as physiological variables and population information; however, concepts provide a much richer source of information on which to base an analysis. For example, a Leximancer analysis for the occurrence of dyspnea and heart attacks could be assessed from a large database of patient records. The results on such an analysis could then be compared against physiological variables and population information to determine if there is a profile for high risk patients with dyspnea. Similarly, the effect of a particular drug or classes of drugs could be initially assessed from patient records without having to encode the information into a format suitable for a database. The major limitation for implementing Leximancer to support medical database mining is the availability of patient case studies and raw patient files. The move in the health care system to electronic patient records [10] means that Leximancer has the potential to serve the needs of clinical and medical informatic analysis.

4 Conclusion

Leximancer provides clinicians and medical informatics experts with the ability to analyze large quantities of patient records and/or patient cases. The results of Leximancer analysis of patient information could be used to identify what information should be sorted from clinical databases and may be used to support other quantitative methods if the database contains both natural language and quantitative information. Early Leximancer assessment of patient records is likely to save time in identifying what quantitative information should be included in medical databases. The inclusion of Leximancer concept analysis as part of database mining is likely to identify relationships that could not be detected by other means.

References

1. Suka, M. Ichimura, T. Yoshida K.: Development of Coronary Heart Disease Databases, Proc. The Eighth Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES'2004) 1081-1088
2. Coiera, E.: Maximising the uptake of evidence into clinical practice - An information economics approach, Medical Journal of Australia, (2001) 174, 467-470
3. Webb, R. K., van de Walt, J., Runciman, W. B., Williamson, J. A., Cockings, J., Russell, W. J., & Helps, S.: Which monitor? An analysis of 2000 incident reports. Anaesthesia and Intensive Care, (1993) 21, 529-542
4. Nisbett, R. E., Wilson, T. D.: Telling more than we can know. verbal reports on mental-processes. Psychological Review, (1977) 84 (3), 231-259
5. Beeferman, D., Berger, A., Lafferty, J.: A model of lexical attraction and repulsion. In the Proceedings of the acl-eacl'97 joint conference. (1997). Madrid, Spain.
6. Smith, A. E. Machine mapping of document collections: the leximancer system. In Proceedings of the Fifth Australasian Document Computing Symposium, Sunshine Coast, Australia. (2000)
7. Smith, A. E.: Automatic Extraction of Semantic Networks from Text using Leximancer. Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics - Companion Volume, ACL (2003) pp Demo23-24.
8. Smith, A. E., Humphreys, M. S.: Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping; Behavior Research Methods, (In press, accepted 29 March 2005).
9. Mahler DA, ed. Dyspnea. Mount Kisco, N.Y.: Futura Publishing, 1990.
10. Van Bemmel J. H., van Ginneken A. M., van der Lei J.: A Progress Report on Computer-Based Patient Records in Europe, in Dick R. S., Steen E. B., Detmer D. E. (eds.), The Computer-Based Patients Record: An Essential Technology for Health Care, rev. ed., IOM, National Academy Press, Washington, (1997) 21-43.

Barrier to Transition from Paper-Based to Computer-Based Patient Record: Analysis of Paper-Based Patient Records

Machi Suka and Katsumi Yoshida

Department of Preventive Medicine, St.Marianna Univeristy School of Medicine,
2-16-1, Sugao, Miyamae-ku, Kawasaki 216-8511, Japan
{suka,k2yosida}@marianna-u.ac.jp

Abstract. To facilitate the transition from paper-based patient record (PPR) to computer-based patient record (CPR), engineers should try to improve the usability of CPR system. From the point of view of a physician, we reviewed PPRs written by 8 Japanese physicians. We revealed the characteristics of PPR to find out about the barrier to transition from PPR to CPR. Our findings may be helpful to the engineers who are aiming to develop a CPR system.

1 Introduction

The Japanese Ministry of Health, Labour, and Welfare promotes the introduction of information technologies into health care and encourages the use of computer-based patient record (CPR). CPR has the potential advantages of (1) saving space and work for record-keeping, (2) sharing information among physicians, staffs, and patients, and (3) handling data promptly and efficiently (search, extraction, tabulation, time series, data mining, etc).

Nowadays a number of CPR systems are available in Japan as well as in the USA and the EU. However, the majority of clinics and hospitals still use paper-based patient record (PPR). There are many reasons why clinics and hospitals hesitate about the transition from PPR to CPR. As issues of hardware, (1) CPR costs a large initial investment, (2) the maintenance of CPR requires some expertise, and (3) CPR is not always compatible with other information systems. As issues of software, (4) CPR provides structured forms and limits usable vocabulary, while PPR permits unstructured free text, (5) the time needed to make out a CPR is longer than that for a PPR, and (6) most physicians feel it troublesome to operate a CPR system. The use of CPR is likely to disturb the work flow and decrease the efficiency of work.

To facilitate the transition from PPR to CPR, engineers should try to improve the usability of CPR system. As the first step, they need to have a better understanding about the gap between PPR and CPR. However, there are few reports that provide information about PPR. From the point of view of a physician, we reviewed PPRs written by 8 Japanese physicians. We revealed the characteristics of PPR to find out about the barrier to transition from PPR to CPR.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1239–1245, 2005.

2 Basic Structure of Patient Record

Physicians are supposed to keep a record whenever they examine a patient. A patient record consists of 4 sections: Subjective (S) that includes symptoms, lifestyle (smoking, drinking, diet/appetite, physical activity, sleep, etc), medication, past history, family history, drug allergy; Objective (O) that includes height, weight, blood pressure, body temperature, physical findings (inspection, auscultation, and manipulation), and laboratory findings; Assessment (A) that includes assessment and diagnosis; and Plan (P) that includes examination, perception, treatment, and next consultation date.

3 Analysis of Paper-Based Patient Records

3.1 Preparation of PPRs

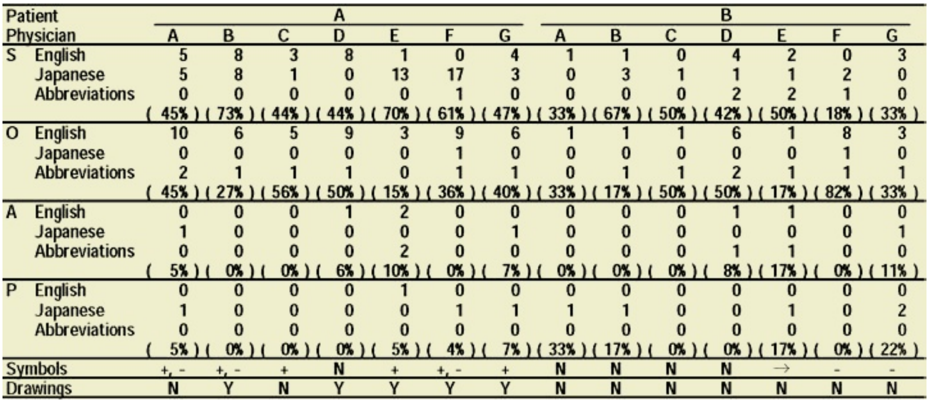
We enrolled 8 Japanese physicians who had been working at clinics and hospitals for more than 10 years. They were asked to examine 2 dummy patients and make out a PPR each patient. The patients were (A) a 35-year-old woman who had a fever (about 38 degrees) and (B) a 65-year-old man who had taken medicine for hypertension and hypercholesterolemia.

3.2 Analysis of PPRs

We reviewed 16 PPRs, 2 PPRs each physician. There were 2 illegible PPRs (written by 1 physician). We therefore analyzed 14 legible PPRs (written by 7 physicians).

Structure. Table 1 shows the distribution of words, symbols, and drawings in the 14 PPRs (except for prescription). In almost all the PPRs, Japanese and English were mixed together and some abbreviations were included in spots. The percentage of English varied according to physician. Symbols and drawings were found in 9 and 5 PPRs, respectively.

Table 1. Distribution of words, symbols, and drawings in the 14 PPRs. Values indicate the number of words. Y means presense. N means absense.



Patient		A							B						
Physician		A	B	C	D	E	F	G	A	B	C	D	E	F	G
S	English	5	8	3	8	1	0	4	1	1	0	4	2	0	3
	Japanese	5	8	1	0	13	17	3	0	3	1	1	1	2	0
	Abbreviations	0	0	0	0	0	1	0	0	0	0	2	2	1	0
		(45%)	(73%)	(44%)	(44%)	(70%)	(61%)	(47%)	(33%)	(67%)	(50%)	(42%)	(50%)	(18%)	(33%)
O	English	10	6	5	9	3	9	6	1	1	1	6	1	8	3
	Japanese	0	0	0	0	0	1	0	0	0	0	0	0	1	0
	Abbreviations	2	1	1	1	0	1	1	0	1	1	2	1	1	1
		(45%)	(27%)	(56%)	(50%)	(15%)	(36%)	(40%)	(33%)	(17%)	(50%)	(50%)	(17%)	(82%)	(33%)
A	English	0	0	0	1	2	0	0	0	0	0	1	1	0	0
	Japanese	1	0	0	0	0	0	1	0	0	0	0	0	0	1
	Abbreviations	0	0	0	0	2	0	0	0	0	0	1	1	0	0
		(5%)	(0%)	(0%)	(6%)	(10%)	(0%)	(7%)	(0%)	(0%)	(0%)	(8%)	(17%)	(0%)	(11%)
P	English	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	Japanese	1	0	0	0	0	1	1	1	1	0	0	1	0	2
	Abbreviations	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		(5%)	(0%)	(0%)	(0%)	(5%)	(4%)	(7%)	(33%)	(17%)	(0%)	(0%)	(17%)	(0%)	(22%)
Symbols		+, -	+, -	+	N	+	+, -	+	N	N	N	N	→	-	-
Drawings		N	Y	N	Y	Y	Y	Y	N	N	N	N	N	N	N

The PPRs consisted chiefly of S and O. English and abbreviations were frequent especially in O, where the physicians had a rather small vocabulary. For native Japanese speakers, Japanese is easier to understand than English, but the writing in Japanese takes longer than that in English. The physicians needed to keep a record in the shortest time. They therefore used English and abbreviations without hesitation.

All of the abbreviations were regarded as jargon: CC=chief complaint; PH=past history; FH=family history; BT (or KT)=body temperature; BP=blood pressure; PR=pulse rate; LN=lymph node; n.c.=no change; n.p.=no problem (or not particular); s/o=suspicious observation. The use of abbreviations is likely to help save time and work for writing but cause interference in information sharing.

Contents. As described in Section 2, a patient record includes many items. Table 2 shows the items in the 14 PPRs (except for prescription). The recorded items were different by patients (i.e. intra-user variance) and physicians (i.e. inter-user variance). The variance seems to be larger in S and O than in A and P.

The items were not always arranged in order. The physicians often added some words to make up for a deficiency and/or to supplement further details. Nonetheless half of the PPRs covered less than 50 % of the items that should be recorded for the patient (Cover % in Table 2). A and P are small but meaningful sections, where the decision of the physician is found. However, not all of the physicians recorded assessment or diagnosis or plans.

Notice that the physicians described the same thing using different words. For example, “condition good”, “doing well”, “no change”, “n.c.”, and “tokuni kawari nashi (no particular change in Japanese)” for Patient B had the same meaning, that is, the patient feels well and he (or she) looks in a good condition. When dating symptoms, 7 physicians counted from the consultation day like “the day before yesterday” and 1 physician specified the date like “1/10 (January 10 in Japanese)”.

As an issue peculiar to Japanese, there were equivocal expressions. For example, “tokuni kawari nashi (no particular change in Japanese)” is used in various situations. If the patient described no symptom in the last consultation, the word means that the patient feels well and he (or she) looks in a good condition. Meanwhile if the patient suffered from a headache in the last consultation, the word means that the patient has not recovered from the headache.

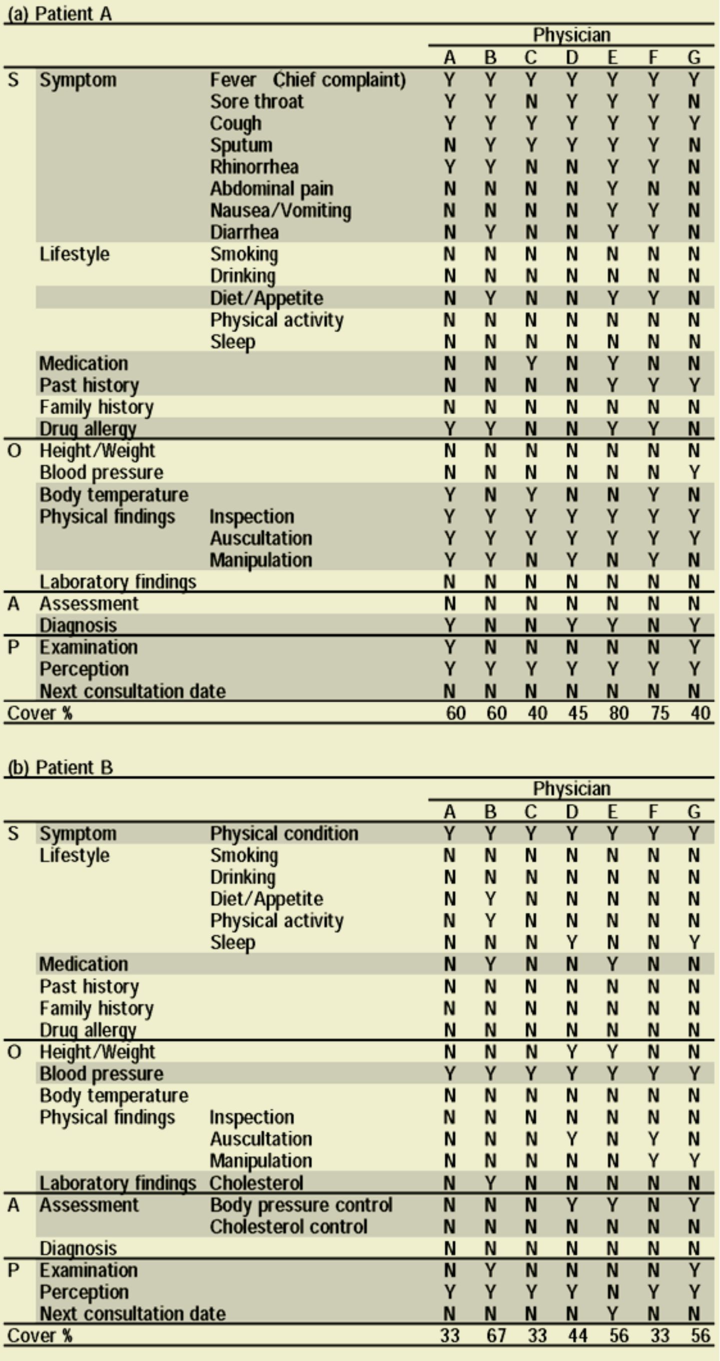
The characteristics of PPR are summarized as follows. (1) illegibility, (2) mixture of languages, (3) use of abbreviations, (4) use of symbols, (5) use of drawings, (6) inexhaustiveness, (7) disregard of order, (8) user-specific vocabulary, and (9) equivocal expressions.

4 Transition from Paper-Based to Computer-Based Patient Record

4.1 State of CPR

Patient records have a comprehensive purpose: to recall observations, to inform others, to instruct students, to gain knowledge, to monitor performance, and to justify interventions [3,4]. Physicians are supposed to keep a record as it meets the purpose in every situation. PPR has the advantage of flexibility, while the illegible and

Table 2. Items in the 14 PPRs. Shadow areas indicate the items that should be recorded for the patient. Y means presense. N means absense.



(a) Patient A

			Physician						
			A	B	C	D	E	F	G
S	**Symptom**	**Fever Chief complaint)**	Y	Y	Y	Y	Y	Y	Y
		Sore throat	Y	Y	N	Y	Y	Y	N
		Cough	Y	Y	Y	Y	Y	Y	Y
		Sputum	N	Y	Y	Y	Y	Y	N
		Rhinorrhea	Y	Y	N	N	Y	Y	N
		Abdominal pain	N	N	N	N	Y	N	N
		Nausea/Vomiting	N	N	N	N	Y	Y	N
		Diarrhea	N	Y	N	N	Y	Y	N
	Lifestyle	Smoking	N	N	N	N	N	N	N
		Drinking	N	N	N	N	N	N	N
		Diet/Appetite	N	Y	N	N	Y	Y	N
		Physical activity	N	N	N	N	N	N	N
		Sleep	N	N	N	N	N	N	N
	Medication		N	N	Y	N	Y	N	N
	Past history		N	N	N	N	Y	Y	Y
	Family history		N	N	N	N	N	N	N
	Drug allergy		Y	Y	N	N	Y	Y	N
O	Height/Weight		N	N	N	N	N	N	N
	Blood pressure		N	N	N	N	N	N	Y
	Body temperature		Y	N	Y	N	N	Y	N
	Physical findings	**Inspection**	Y	Y	Y	Y	Y	Y	Y
		Auscultation	Y	Y	Y	Y	Y	Y	Y
		Manipulation	Y	Y	N	Y	N	Y	N
	Laboratory findings		N	N	N	N	N	N	N
A	Assessment		N	N	N	N	N	N	N
	Diagnosis		Y	N	N	Y	Y	N	Y
P	**Examination**		Y	N	N	N	N	N	Y
	Perception		Y	Y	Y	Y	Y	Y	Y
	Next consultation date		N	N	N	N	N	N	N
Cover %			60	60	40	45	80	75	40

(b) Patient B

			Physician						
			A	B	C	D	E	F	G
S	**Symptom**	**Physical condition**	Y	Y	Y	Y	Y	Y	Y
	Lifestyle	Smoking	N	N	N	N	N	N	N
		Drinking	N	N	N	N	N	N	N
		Diet/Appetite	N	Y	N	N	N	N	N
		Physical activity	N	Y	N	N	N	N	N
		Sleep	N	N	N	Y	N	N	Y
	Medication		N	Y	N	N	Y	N	N
	Past history		N	N	N	N	N	N	N
	Family history		N	N	N	N	N	N	N
	Drug allergy		N	N	N	N	N	N	N
O	Height/Weight		N	N	N	Y	Y	N	N
	Blood pressure		Y	Y	Y	Y	Y	Y	Y
	Body temperature		N	N	N	N	N	N	N
	Physical findings	Inspection	N	N	N	N	N	N	N
		Auscultation	N	N	N	Y	N	Y	N
		Manipulation	N	N	N	N	N	Y	Y
	Laboratory findings	**Cholesterol**	N	Y	N	N	N	N	N
A	**Assessment**	**Body pressure control**	N	N	N	Y	Y	N	Y
		Cholesterol control	N	N	N	N	N	N	N
	Diagnosis		N	N	N	N	N	N	N
P	**Examination**		N	Y	N	N	N	N	Y
	Perception		Y	Y	Y	Y	N	Y	Y
	Next consultation date		N	N	N	N	Y	N	N
Cover %			33	67	33	44	56	33	56

unstructured records are hardly adequate to support information sharing and decision making. There are great hopes that the transition from PPR to CPR will bring about a solution to the problem and improve the quality of clinical practice [1,2,5,6].

However, CPR systems still leave room for improvement: (1) data entry should be simple and easy, (2) relevant information, which is often stored in different databases, should be simultaneously displayed on the screen, (3) the system should authenticate more than one user, (4) the system should provide substantial data security, and (5) the system should have the measures to prevent falsification of records. Above all, the usability of CPR system is a matter of the greatest concern to physicians. The development of user-friendly data entry may be the key to the transition from PPR to CPR.

4.2 User-Friendly Data Entry

Most CPR systems operate on a laptop or desktop computer, which is provided with point-and-click or pick-list interfaces (i.e. templates) for data entry. Nowadays tablet personal computers are available to the general public. The application of graphic tablet technology to CPR system has started to attract attention. The data entry by pen, which gives a feeling of hand-writing, may contribute to closing the gap between PPR and CPR. Unfortunately, hand-writing recognition is far from complete at the moment.

Templates prescribe what items should be recorded for the patient. After displaying a template on the screen, the user checks the applicable box or clicks on the appropriate category. The template entry has little flexibility but helps avoid omission and ambiguous information. Moreover, the use of CPR is likely to cause changes in the information gathering and reasoning strategies and lead the user (physician) to logical thinking [1,2]. There are great hopes that good templates will improve not only the quality of patient records but also the quality of clinical practice.

Templates should be consistent with clinical guidelines and expert opinions [2]. As shown in Table 2, the items to be recorded are considerably different by patient. Templates therefore should change their items according to (1) whether he (she) is a new or repeat patient and (2) what symptoms he (she) describes. A variety of templates should be prepared, each for a different kind of patient.

As shown in Fig.1, a template can be expressed as a hierarchy diagram. Physicians are supposed to gather detailed information only when they find some problem item.

For example, if the patient has a cough, the physician will ask when the cough started (date), what time the cough becomes worse (time), and whether the cough is dry or wet (sputum). Meanwhile if the patient does not have a cough, the physician will leave off asking about cough. Not all of the items are equivalent in clinical meaning. To save time and work for data entry, a template should include exclusively of minimum necessary items. Moreover, the items should be arranged in order of hierarchy: the corresponding subordinate items are displayed or concealed, according as the user checks the box of superordinate item. The development of interactive graphical user interface may lead to user-friendly data entry.

If possible, templates should permit modification to suite the needs or preferences of individual user. The user-specific templates may relieve the reluctance to use CPR.

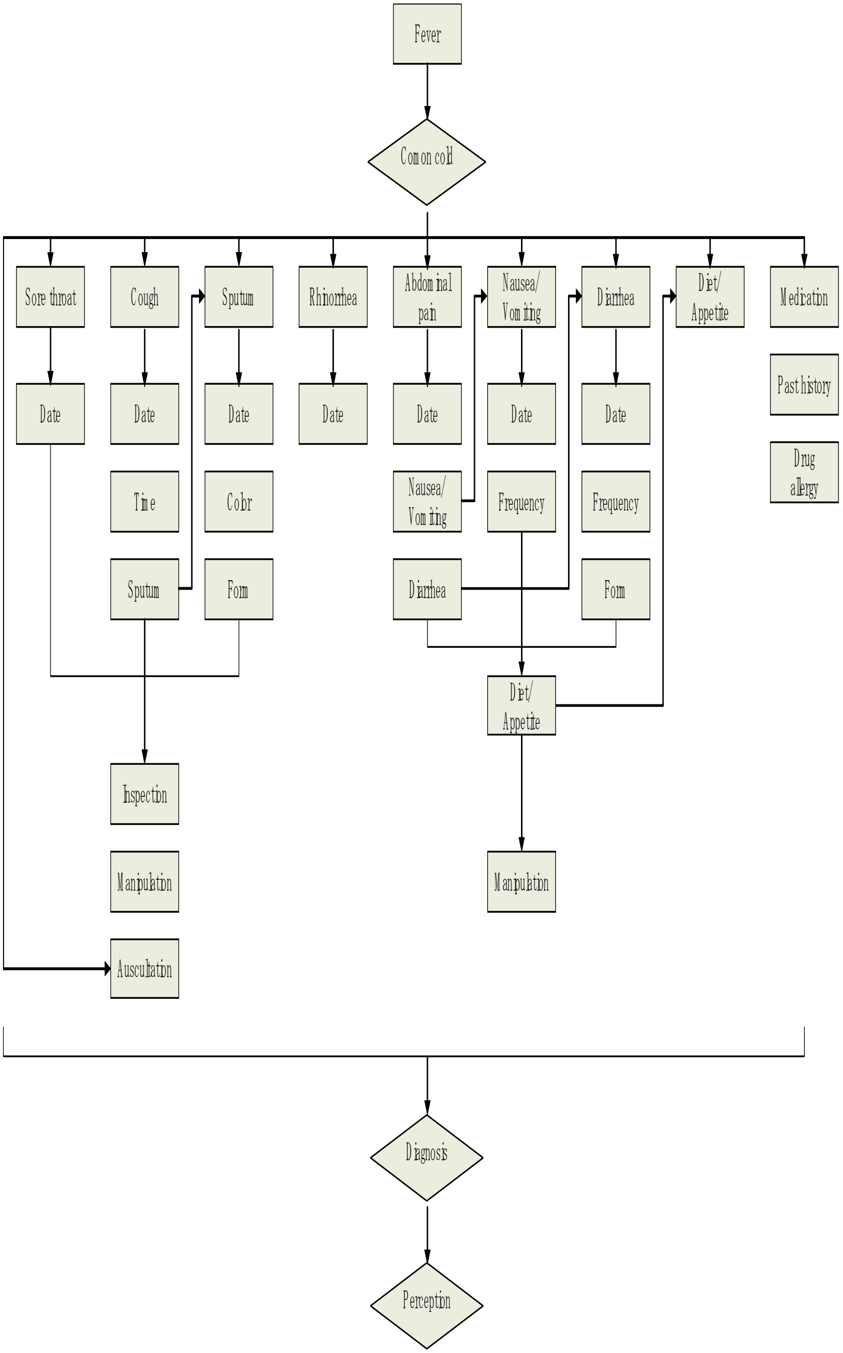


Fig. 1. Hierarchy diagram for the patient who has a fever (Patient A in Section 3)

5 Hopes for Engineers

The transition from PPR to CPR is imperative to cover the shortcomings of PPR. However, not all of the CPR systems have been developed from the point of view of a physician. It is important to know what physicians do and what they really hope for CPR systems. The discussions between engineers and physicians will pave the way for improving the usability of CPR system, which will contribute to removing the barrier to the transition from PPR to CPR.

Acknowledgements

We would like to thank all the anonymous physicians for their kind cooperation. We are grateful to Dr. Takumi Ichimura for his useful advice.

References

1. Patel V.L., Kushniruk A.W., Yang S., Yale J.: Impact of a computer-based patient record system on data collection, knowledge organization, and reasoning. J. Am. Med. Inform. Assoc. 7 (2000), 569-585
2. Elson R.B., Connelly D.P.: Computerized patient records in primary care: Their role in mediating guideline-driven physician behavior change. Arch. Fam. Med. 4 (1995) 698-705
3. Reiser S.J.: The clinical record in medicine. Part 1. Learning from cases. Ann. Intern. Med. 114 (1991) 902-907
4. Reiser S.J.: The clinical record in medicine. Part 2. Reforming content and purpose. Ann. Intern. Med. 114 (1991) 980-985
5. Tang P.C., LaRosa M.P., Gorden S.M.: Use of computer-based records, completeness of documentation, and appropriateness of documented clinical decisions. J. Am. Med. Inform. Assoc. 6 (1999) 245-251
6. Hippisley-Cox J., Pringle M., Cater R., Wynn A., Hammersley V., Coupland C., Hapgood R., Horsfield P., Teasdale S., Johnson C.: The electronic patient record in primary care - regression or progression? A cross sectional study. BMJ 326 (2003) 1439-1443

Preprocessing for Extracting Information from Medical Record to Add XML Tags

Yoshiaki Kurosawa[1], Akira Hara[1], Machi Suka[2], Takumi Ichimura[1]

[1] Faculty of Information Sciences, Hiroshima City University,
3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194, Japan
{kurosawa,hara,ichimura}@its.hiroshima-cu.ac.jp
[2] Department of Preventive Medicine, St. Marianna University School of Medicine,
2-16-1, Sugao, Miyamae-ku, Kawasaki 216-8511 Japan
suka@marianna-u.ac.jp

Abstract. The purpose of this paper is to propose preprocessing procedures from handwritten medical records, of which words are omitted, to translate contents of the records into formatted data such as XML data. From the viewpoint, we performed an experiment, which physicians freely described patients' problems to reveal the characteristics in the handwritten record. As a result, we obtained five characteristics. Based on these characteristics, we considered eight preprocessing steps. Then, we confirmed that our proposed preprocessing worked well from a computational experiment although not all sub-procedures are effective due to lack of a large amount of corpora.

1 Introduction

Recently, electric medical record has been introduced into medical care scene because the record is convenient for searching appropriate medicines, choosing one of them, and computing the mark of them. As to information processing scene as well as medical one, the data is convenient for dealing with the included data because of the obviously unique structure, and it is easy to be able to translate the structuring data into XML (eXtensible Markup Language) data, e.g., Medical Markup Language: MML (MedXML consortium, 2003). Using such data, we can easily let systems find new hidden structures and unknown rules: diagnostic rules (Hara et al., 2004, 2005).

This convenience of the record, however, may not be for physicians, at least, in terms of their descriptions of patients' diseases. Certainly, From the viewpoint of particular medial acts, such as the selection of necessary medicines, and our interest to find new diagnostic rules, the record is convenient because its input is formatted, but in case of the description of patients' diseases, it is difficult to adjust their symptoms to predetermined style, because the symptoms differ from each other, and the description of them depends on physicians' preferences. Therefore, it is obvious that physicians need no restrictions to describe them in consideration of existing studies for some pen-based computer system, which allow them to describe all or part of the electric medical record with no or less restrictions (Igarashi et al., 2000).

We consider physicians prefer the description without any restrictions. For example, see Fig.1 in Section 3 freely handwritten by physicians, which we will explain later in detail. However, the description makes various uses of its contents, e.g., translation into XML data, difficult in the computer science, because the words in the de-

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1246–1252, 2005.

scription are omitted, the contents are ambiguous, and our system clearly does not understand what is described. For this reason, it is important that we precisely analyze such ambiguous description in medical record.

Thus, in this paper, we consider necessary preprocessing procedures as a basis of translation, in other word, how to analyze physician's description in medical record to translate the description including his/her preferences into formatted data such as XML. In addition, we perform an experiment that physicians practically describe patients' problems and make an analysis of the characteristics of their description.

2 SOAP Format in Medical Record

We briefly explain medical record from the viewpoint of the SOAP format.

2.1 SOAP Format

The description based on the Problem-Oriented Medical Record (POMR) is recommended when recording in medical care scene (cf. Toritsubyouinn Shinryourokutou Kisaikenntouiinnkai, 2001). In this POMR description, physicians need to describe patients' problems, which the patients encounter one or more problems such as fever, cough, and sore throat, according to the SOAP format. SOAP stands for Subjective, Objective, Assessment, and Plan as described below.

- Subjective: One or more patients' problems are described.
- Objective: Physical examination findings and test results are described.
- Assessment: Physicians' notes including diagnoses are described.
- Plan: One or more treatment plans are described.

2.2 Electric Medical Record from the Viewpoint of the SOAP Format

As well as being attempted to standardize as XML structures, e.g., "progress course information module" (MedXML consortium, 2003), the SOAP format is implemented in case of the description of electric medical record (cf. Poon and Fagan, 1994; Igarashi et al., 2000). Their implemented systems are divided into two different types. The first one notes on making use of the data according to this format, and the second one notes on easiness to input physicians' ideas without any restrictions.

First, Poon and Fagan adopt the format and propose the interface using pen-based system, which physicians can select the most appropriate choice from the list of choices and circle it. For example, the list "Constitutional" have 10 choices such as fever, chills and sweats, and we can select one or more symptoms, e.g., fever. Therefore, their system can translate this selection into formatted data.

Second, Igarashi et al. also adopt the SOAP format and propose the pen-based interface which physicians can describe their findings, as they prefer. Thus, this system has no restrictions when recording. On the other hand, the study of Poon and Fagan has restrictions that physicians must select from the list [1] and cannot select any non-

[1] Their system allows physicians to describe handwritten notes in free-text form. However, we consider it has restrictions because their main concern is not to describe patients' problems freely but to select choices using a pen-based interface

existent choices. From this standpoint, Igarashi et al.'s system is useful. However, the system includes a certain problem that we cannot make use of the input easily because it deals with the input as not text data but vector data. For this reason, we need to solve the problem.

3 Analysis of Medical Record

The SOAP format is used when physicians describe patients' problems in medical record as we mentioned in previous section. However, every physician does not describe his/her medical record according to this classification related to this format, because the classification is only recommended and is not indispensable for the description.

Then, how do physicians record their notes? In this section, we investigate and analyze the characteristics of physicians' practical descriptions in medical record.

3.1 Experiment

We experimented to obtain sample medical records. We explain the procedure.

Participants: Seven physicians

Patient's problems:

We prepared for two cases: fever and high-blood pressure. In addition, the former was supposed his/her first visit. The latter was supposed his/her revisit.

Procedure:

Each participant examines two patients mentioned above. His/her task is to make patients' records on paper with horizontal ruled line as he/she always does.

Fig.1, as shown in next page, indicates two physicians' sample medical records in case that patient's problem is "fever."

3.2 Analysis of the Characteristics of Medical Record

We explain five characteristics captured from our experiment.

- Virtual Region

  Physicians seemed to use regions without explicit boundaries. For example, the term "Rp," recipe, emerged after indented in Fig.1L (left).

  The SOAP format mentioned in Section 2 was used under another form in Fig.1L; the abbreviations "A" and "P" were described at the same place. On the other hand, Fig.1R(right) illustrated the nonnecessity of the SOAP format even though two physicians described their records without the format.

  This characteristic is important to make various uses of data because we can regard the abbreviations "Rp," "A" and "P" related to the format as one of the positive examples in case of performing supervised learning. For instance, using the expression "possible influenza case" and the other elements such as patients' symptoms in the field "O," our system can learn rules to detect their diseases.
- Illustrations and marks

  There were some illustrations, as shown in Fig.1R. Besides that, both figures included two types of marks related to "plus" and "minus."

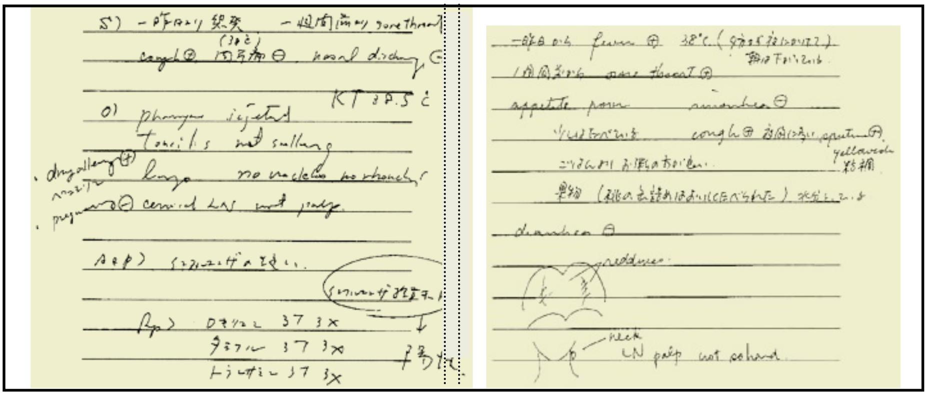

S)
KT 38.5
O) pharynx injected
tonsils not swollen
cervical LN not palp.
A&P)
Rp)
fever ⊕ 38℃
sore throat ⊕
appetite poor
cough ⊕
sputum ⊕
yellowish
diarrhea ⊖
redness
neck
LN palp not palpable

Fig. 1. Two Sample Records

- Multi-Language
  No data was recorded only in Japanese. One was written only in English and the rest (13 data) consisted of Japanese and English, as shown in Fig.1L and Fig.1R. In addition, German emerged as an abbreviation.
- Technical Terms and Abbreviations
  We found many technical terms such as the name of medicine and disease. Furthermore, we also found various abbreviations such as BP (Blood Pressure), BT (Body Temperature) and KT (Body Temperature in German).
- Ellipses
  Not a few ellipses emerged. For example, an expression "cough" was described without another expression. Moreover, another expression "*Syoutyuu*, 1-2 *hai* (one or two glasses of a kind of liquor)" was seen without a verb *nomu* (drink).

In order to understand ambiguous handwritten data described by physicians and translated it into formatted date such as XML data, we must take these characteristics into account and implement our system. In next section, we explain a necessary processing flow of our preprocessing in consideration of this subsection.

4 Processing Flow of Preprocessing

As we explained in Section 3, the characteristics of medical record make it difficult to analyze the record. In this section, we consider a series of procedures to solve this difficulty. Eight steps are required at least.

- Step 1: Region detection and OCR
  This procedure is used to solve the problem "Virtual Region." Certainly, we cannot detect the virtual regions imaged by physicians because there are no explicit borders, but we consider that calculating the two-dimensional and temporal distance between words and clustering them, we can solve the problem. Above all, the temporal distance may be effective because physicians should describe the records continuously as far as the same topic continues. However, due to lack of knowledge as to such ergonomic respect, we manually divide the records into some appropriate regions in this paper.

We also manually translate the handwritten medical records into text files due to low precision in our OCR test.

- Step 2: Conversion Marks to Language

  This conversion process is basically simple. We only convert the mark "(+)" into Japanese word "*ari*," which means "(the patient) has," and the mark "(-)" into "*nashi*," which means "(the patient) does not have." Therefore, the phrase "*seki* (+)" is converted into "*seki ari*," which means "the patient have a cough."

  In addition, we ignore undefined marks and any figures.

- Step 3: Translation

  This is concerned with the characteristics "Multi-Language." We require translation into Japanese to perform morphological analysis at latter stage. In addition, this translation means simple replacement of an English word with a Japanese word because the sentence is simple in medical record. For example, the description "appetite poor" is translated into "*syokuyoku toboshii*."

- Step 4: Morphological Analysis

  We adopt a morphological analyzer ChaSen (Matsumoto et al., 2000), of which morphemes, i.e., technical terms, are added in its dictionary beforehand.

- Step 5: Automatic Registration of Technical Terms

  In general, the appearance of unregistered morphemes and abbreviations occurs morphological errors. We can correct the errors in case of well-known errors using correction rules (Kurosawa et al., 2003, 2005a). However, if similar kinds of errors have never occurred, we need to find the errors manually and then, we must make rules to correct them. It is difficult that we precisely perform first task without mistakes. For this reason, we adopt an error measure to detect the errors automatically (Kurosawa et al., 2005b). Using this procedure, we attempt to solve the problem related to "Technical Terms and Abbreviations."

- Step 6: Correction of Morphological Errors

  As we explained in Step 5, we use the rules for correction and modify morphological errors automatically (Kurosawa et al., 2003, 2005a).

- Step 7: Ellipsis Resolution

  This step resolves ellipses because the existence of implicit words and phrases make it difficult to translate the description into formatted data.

  For example, the noun "cough," as mentioned in "Ellipses," appears by itself. We can find the verb phrase ellipsis because one noun usually does not consist of a sentence without verbs. Such heuristics contributes the findings of the ellipses. However, our system cannot detect what morphemes are appropriate to these ellipses although we understand that the verb "have" is necessary for "cough."

  Of course, if we can observe the frequencies of co-occurrence between two words, e.g., "cough" and "have" from a large amount of medical corpora, our system may detect the verb related to the ellipsis. Unfortunately, we cannot deal with this type of resolution because we do not have the corpora.

- Step 8: Context Analysis

  This step deals with certain type of ellipses that we can detect the nonexistent morphemes using context although we cannot detect them in Step 7. The term "context" indicates various types of useful information in various regions: each sentence as a region, each section detected in Step 1, and so on.

For example, the expression "*Syoutyuu*, 1-2 *hai*," as we mentioned in "Ellipses," cannot make us detect the necessary verb for the same reason as we explained in Step 7. However, in case that some preceding sentences such as "I drink alcohol every day." emerge before the word "*Syoutyuu*" and, the word "alcohol" has something to do with the word "*Syoutyuu*," we consider the expression "*Syoutyuu*, 1-2 *hai*" requires the verb (drink) included in the preceding sentence. This relation between words is understood using a concept dictionary: *GoiTaikei* (Ikehara et al., 1997). Thus, we can detect the nonexistent morphemes.

4.1 Experimental Evaluation

Using 14 data mentioned in Section 3, we performed a computational experiment and confirmed that our processing flow worked well. For example, all description "*seki*," "*seki ari*," "*seki* (+)" and "cough (+)" were interpreted as a unique result "*seki ari*" ("have a cough"). Thus, we could describe it using XML data as follows:

```
<mmlPc:ProgressCourseModule>
      <mmlPc:FreeExpression>
            seki  ari <xhtml:br/> …
      </mmlPc:FreeExpression>
</mmlPc:ProgressCourseModule>
```

However, as we mentioned previously, we do not deal with all steps due to lack of related data and corpora; Step 1, 5 and 7 are useless in this paper. Moreover, in Step 8, only a few samples can be analyzed because of the small size of data although we consider Step 8 is important and useful. On the grounds of these matters, we merely obtained almost the same knowledge, e.g., related to various morphological errors, which we have already obtained from experiments (Kurosawa et al., 2003, 2005a).

5 Conclusion and Future Works

We performed an experiment which physicians described patients' problems and analyzed the data from the experiment to reveal the characteristics in handwritten medical record. In consideration of this analysis, we proposed eight preprocessing steps for extracting information from electronic medical record, of which words were omitted and contents are ambiguous.

As a result of our computational experiment, we confirmed that our proposed preprocessing functioned well. However, due to lack of some ergonomic experimental data and medical corpora, our evaluation is less effective. Therefore, we need two future works as follows: an ergonomic experiment to divide medical records into some appropriate regions (Step 1) and development of a large amount of corpora (Step 5 and Step 7). The former is required to observe physicians' practical writing behavior, in particular, physicians' temporal factor such as writing time and interval time for each section of the SOAP format. The latter is required to perform two computational experiments, automatic detection of morphological errors (Step 5) and ellipsis resolution (Step 7), after development of corpora, because these experimental results depend on the advantage and disadvantage of the corpora.

Through these consideration, we need to totally consider whether our proposed procedures are effective or not. Then, we will translate our output analyzed by our procedures into certain formatted data, and also attempt to find unknown diagnostic rules.

References

1. Hara, A. et al. (to appear in 2005), "Extraction of rules from coronary heart disease database using automatically defined groups," International Journal of Manufacturing.
2. Hara, A. et al. (2004), "Extraction of rules from coronary heart disease database using automatically defined groups," In *Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, pp.1089-1096.
3. Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., Oyama, Y., and Hayashi, Y. (1997), "*GoiTaikei - A Japanese lexicon*," Iwanami Shoten (in Japanese).
4. Igarashi, T., Ashihara, T., Nagata, S., Takada, M., and Nakazawa, K. (2000), "A pen-based interface for electronic medical recording systems: toward stress-free experience for doctors," In *Proceedings of the 20th Joint Conference on Medical Informatics* (in Japanese).
5. Kurosawa, Y., Ichimura, T., and Aizawa, T. (2003), "A description method of syntactic rules on Japanese filmscript," In *Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, pp.446-453.
6. Kurosawa, Y., Ichimura, T., and Aizawa, T. (2005), "A description method of syntactic rules on filmscripts," *Journal of Natural Language Processing*, 12(2), pp.25-62 (in Japanese).
7. Kurosawa, Y., Sakamoto, Y., Ichimura, T., and Aizawa, T., "An error measure for Japanese morphological analysis using similarity measures," (submitted to KES2005).
8. Matsumoto, Y. et al. (2000), "Morphological analysis system ChaSen version 2.2.1 manual," http://chasen.aist-nara.ac.jp/.
9. MedXML consortium. (2003), "MML (Medical Markup Language) Version 3.0 Specification," http://www.medxml.net/E_mml30/.
10. Toritsubyouin Shinryourokutou Kisaikenntouiinnkai. (2001), "*Shinnryourokutou Kisai Manual* [The metropolitan hospital manual on how to write medical records]," Bureau of Public Health, http://www.byouin.metro.tokyo.jp/osirase/hokoku/sinryoroku.pdf (in Japanese).
11. Poon, A. D. and Fagan, L. M. (1994), "PEN-Ivory: the design and evaluation of a pen-based computer system for structured data entry," In *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, pp.447-451.

A Scheduling Method of Data Transmission in the Internet Communication by Recurrent Neural Network

Norio Ozaki[1] and Takumi Ichimura[2]

[1] Department of IT Business, Chugai Technos Inc.,
9-12, Yokogawa-shinmachi, Nishi-ku, Hiroshima 733-0013, Japan
n.ozaki@chugai-tec.co.jp
[2] Faculty of Information Sciences, Hiroshima City University,
3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194, Japan
ichimura@its.hiroshima-cu.ac.jp

Abstract. We developed the Remote Monitoring Intelligent System (RMIS) of Local Area Network (LAN) to realize the secure Internet environment. The RMIS has various agents which work to exchange the information related to computers condition with the network environment. We shall be able to control the quality and quantity in the network communication by their agents. In this paper, we propose the scheduling method of data transmission by Recurrent Neural Networks(RNNs) to avoid traffic jam in the network. Especially, our proposed RNN enables to expect the condition of network at each network device instead of time series data. In order to verify the effectiveness of our proposed method, we report the examination results.

1 Introduction

Recently, the Internet oriented information society has been enriched our intellectual lifestyle with various communication tools such as ubiquitous computing, mobile communications and so on. However, we may meet the unexpected information such as invading the computer or tapping the authentication system, and then working destructive operations and trespassing upon our privacy[1].

In order to impair our damage, the network security technologies have been developed from the various viewpoints. The firewall is known that the network communication device connected from the network to other networks and we can give a limitation of the use for unnecessary TCP/IP ports. For the control of TCP/IP ports, we should always be aware of the vulnerabilities in the operating software applications in the corresponding TCP/IP ports. In addition, the receiving the e-mail and the retrieval results of database through web site may be attached the virus. When we consider the use of a clinic Local Area Network(LAN) connected the Internet, we will protect personal information against destructing, trespassing, or virus.

In this paper, we developed the Remote Monitoring Intelligent System (RMIS) of LAN to realize the secure Internet environment. The RMIS has a Management Server(MS) and the plural Firewall communication devices of each LAN described in the chapter 2. The system includes various agents to exchange the information related to computers condition with the network environment. The agent in the MS is always

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1253–1259, 2005.

observing the registered Firewalls. Each agent in the Firewall exchanges the agents in the servers of the corresponding network and provides the self-repairing functions to the servers.

Moreover, this paper proposed the scheduling method of Recurrent Neural Network (RNN) to perform an effective data transmission between a MS and firewalls. The usual RNN is estimated by time series data, but this method can evaluate whether the data transmission can reach within expected time through various network devices such as routers. In order to verify the effectiveness of the proposed method, we report learning results of the RMIS of Secure LAN.

2 Remote Monitoring Intelligent System of Secure LAN

In order to avoid damages by attacking or spoofing, we developed the RMIS of LAN where some agents detect intrusion and communicate its information with each other.

This chapter describes the architecture of the RMIS and the functions of their agents.

2.1 The System Overview

Fig.1 shows an overview of the RMIS of secure LAN. The MS monitors the network condition of LAN_1 to LAN_m which are connected to the Internet by our developed firewall, respectively.

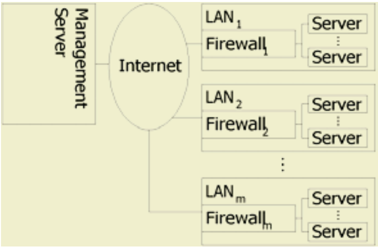


Fig. 1. An overview of RMIS

2.2 The Function of Agents

There are three kinds of agents in the RMIS; MS Agent, Firewall Agent and Server Agent in the LAN. They aim to help the good condition of data transmission in the Internet and to keep the security of each LAN cooperatively by the exchange of the security information such as attacking or spoofing. Each agent fulfills its function as follows.

MS Agent operates

- to learn the communication records while distributing packaged software and to make a schedule of transmission to firewalls by the RNN through network devices.
- to estimate if the software can be transmitted within the expected time.
- to actually send the entire software.

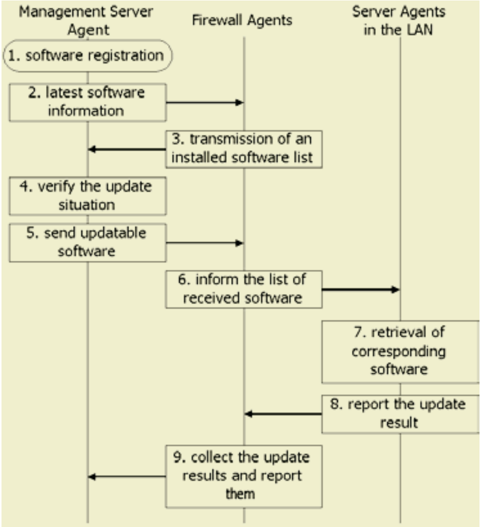

Management Server Agent
Firewall Agents
Server Agents in the LAN
1. software registration
2. latest software information
3. transmission of an installed software list
4. verify the update situation
5. send updatable software
6. inform the list of received software
7. retrieval of corresponding software
8. report the update result
9. collect the update results and report them

Fig. 2. A flow chart of software update procedure

- to record the transmitted softwares to firewall.
- to inform that finishes sending software to each firewall.
- to receive a list of name and version of the resisted software in all servers worked in the LAN through each firewall.

Firewall Agent operates

- to store the received software from the MS temporally.
- to make a list of name and version of completely installed software in the servers in the corresponding LAN and communicate it to the MS.
- to record name and version of the above list in its own database.

Server Agent in the LAN operates

- to receive the update software from the firewall.
- to update software automatically.
- to notice the list of name and version of the updated software to the Firewall Agent.

2.3 The Flow Chart of Software Updates

If we found the vulnerability such as the security holes, errors, or bugs in the operating software, we must repair the weak points immediately. The RMIS can provide the automatically update service for the corresponding packaged software as shown in Fig.2.

2.4 The Scanning Vulnerability of Software

Firewall Agent will find vulnerability of server's software to scan security hole after update procedure. Fig.3 shows an overview of scanning vulnerability.

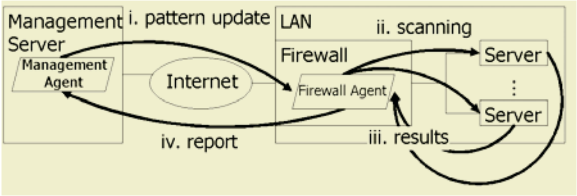


Fig. 3. An overview of scanning vulnerability

3 Recurrent NN Through Network Devices

The general RNN is designed to expect a condition at time TTL (Time To Live) using the time series data. The Back Propagation Through Time (BPTT) and the Epochwise BPTT (EBPTT) models are well known to resolve such problems [2][3].

However, the communication speed through some network devices is changed according to the amount of IP packets. Then, it is difficult to measure the time series data of TTL by using the network commands such as “ping” or “traceroute”.

We assume that we may observe listen to the intermediary nodes between the MS and the Firewall and then, the RNN is trained by the results of the observations. This section gives a shortly explanation of learning and expectation of our proposed RNN.

3.1 Learning of Data Communication

The data transmission between MS and Firewall passes through two or more network devices as shown in Fig.4. The H denotes the number of “hops” from MS to Firewall. The proposed method can learn the amount of network traffic through all network devices by using “hop” in the “ping” or “traceroute”.

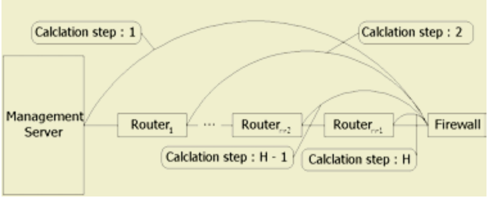


Fig. 4. A data transmission between MS and Firewall

Fig.5 shows an overview of our proposed method. Our method is different from general EBPTT with respect to the insertion of hidden neurons. In Fig.5, x_i $(1 \le i \le 10)$ and y_j $(1 \le j \le 2)$ denotes the set of input and output signals, respectively. The $d_j[h]$ $(1 \le h \le H)$ denotes the desirable signals at the intermediate node h. The output of y_1 in calculation step h turns into an input in $h+1$. The input and output neurons are consisted of the specified terms as follows x_i =(x_1 :start time of transmission data, x_2 :day of a week, x_3 :holiday, x_4 :the classification ID of firewalls, x_5 :data size, x_6 :received packets, x_7 :sending packets, x_8 :the number of hops, x_9 :reachable time to the firewall, x_{10} :expected time of transmission) and y_j =(y_1 :end time of transmission, y_2 :a possibility of completely data transmission within expected time).

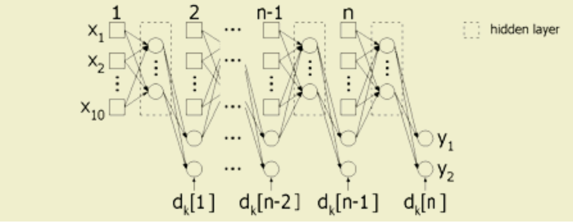


Fig. 5. An overview of our proposed RNN

At h , input neuron x_8 :

$$x_8[h] = n - h + 1 \tag{1}$$

We can obtain an arrival time to the Firewall from the MS or router by using the network commands “ping” and “traceroute”[4]. As a result, x_9 can be estimated from the TTLs. The teaching signal $d_1[h]$:

$$d_1[h] = d_1[0] - x_9[h] \tag{2}$$

where $d_1[0]$ is end time of transmission from MS to Firewall.

The hidden neuron z_k $(1 \le k \le M)$ and the output neuron y_j are calculated by (3) and (4), respectively. The R denotes the number of neurons with feedback connection.

$$z_k[h] = f\left(s_k[h]\right), \quad s_k[h] = \sum_{i=1}^{I}\left(w_{i,k} \cdot x_i[h]\right) + \sum_{j}^{R}\left(v_{j,k} \cdot y_j[h-1]\right) \tag{3}$$

$$y_i[h] = f\left(t_j[h]\right), \quad t_j[h] = \sum_{k=1}^{M}\left(u_{k,j} \cdot z_k[h]\right) \tag{4}$$

where $w_{i,k}$ and $v_{j,k}$, $u_{k,j}$ denotes the weight vector, respectively. The weight are modified by the following equations. The η is a learning rate, and O is the number of output neurons.

$$\Delta u_{k,j} = -\eta \sum_{h=0}^{H}\left(\delta_j[h] \cdot z_k[h]\right), \; \Delta w_{i,k} = -\eta \sum_{h=0}^{H}\left(\varepsilon_k[h] \cdot x_i[h]\right),$$
$$\Delta v_{j,k} = -\eta \sum_{h=0}^{H-1}\left(\varepsilon_k[h+1] \cdot y_j[h]\right), \tag{5}$$

$$\text{where } \varepsilon_k[h] = z_k[h] \cdot (1 - z_k[h]) \cdot \sum_{j=1}^{O}\left(\delta_j[h] \cdot u_{j,k}\right)$$

$$\begin{aligned} &\mathit{if}\ h = H\ \mathit{or}\ R < j \le O \quad && \delta_j[h] = y_j[h] \cdot (1 - y_j[h]) \cdot (y_j[h] - d_j[h]) \\ &\mathit{otherwise} && \delta_j[h] = y_j[h] \cdot (1 - y_j[h]) \cdot \left\{ (y_j[h] - d_j[h]) + \sum_{k=1}^{M}\left(\varepsilon_k[h+1] \cdot v_{j,k}\right) \right\} \end{aligned} \tag{6}$$

The network is trained until the squared error E becomes sufficiently small by BP learning.

$$E = \frac{1}{2}\sum_{p=1}^{P}\sum_{h=1}^{n_p}\sum_{k=1}^{O}\left(d_k^p[h] - y_k^p[h]\right)^2 \tag{7}$$

where P is the number of patterns and n_p is the number of hop in the pattern p .

3.2 Estimating Method of Data Transmission

This research estimates a possibility of completely sending of software data within the expected time. After the learning of the RNN, we can obtain the possibility of completely data transmission within expected time by y_2 . In this simulation, we assume that data transmission can completely implemented if the value is larger than 0.5.

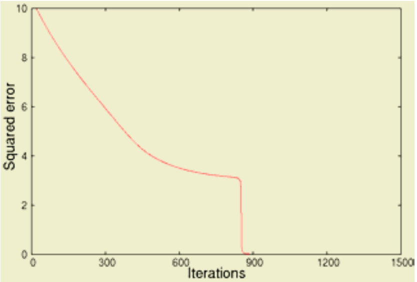


Fig. 6. Error convergence

Table 1. The training data

pattern	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y_1	y_2
1	23:45:10	Tue	0	1	1	60	40	22	47,46,…	10	11	0
2	19:45:10	Wed	0	1	1	60	40	22	47,46,…	10	9	1
p	19:45:10	Wed	0	1	1	60	40	22	47,46,…	10	9	1

3.3 Scheduling Method

In this section, the schedule plan of data transmission to each Firewall is drawn up based on the estimation results as follows.

1) First, the input signals of RNN is given $x_{10} = a$ denoting the usual transmission time as an initial value. In 1), if the estimation result shows the "no reachable" (the y_2 is less than 0.5), MS Agent set the $x_{10} = 2a$ and estimate again.
2) MS Agent estimate against all Firewalls and records the results. The records are sorted in small order of x_{10} .
3) MS Agent transmits the corresponding software actually to each Firewall according to the schedule.

4 Experimental Results

In order to verify the effectiveness of our proposed method, the calculation simulation is implemented. The training data are consisted of 10 ideal records. A part of the training dataset is shown in Table.1. Fig.6 shows the error curve of squared error.

5 Conclusive Discussion

In this paper, we proposed learning and estimating method of RNN through network device in RMIS where MS Agent and Firewall agents are detected the traffic condition in the network to communicate their measurement results each other. MS Agent can make a schedule list of transmission of software data. The security system can monitor the network communication and repair itself if the operating system has vulnerabilities. Therefore, the system is effective to use in a clinic LAN, even if the network is connected to the Internet.

The RTT(Round Trip Time) measured by “ping” or “tracerote” are about several mille-seconds and tends to increase when the number of “hops” increases. However, we may meet that the average RTT is large in spite of the number of “hops” is small. Such a phenomena is caused by the performance of the network device. Such data are eliminated from the dataset because of the delay of ICMP process at the routers. For the statistical elimination, we consider that the abnormal signals are eliminated based on normal distribution of data automatically.

References

1. CERT Coordination Center(2005), http://www.cert.org/
2. R. Williams and D. Zipser: A learning algorithm for continually running fully recurrent neural networks, Neural Computation, Vol.1, pp.270-280(1989).
3. R. Williams and D. Zipser: Gradient-based learning algorithms for recurrent networks and their computational complexxity, in Y. Chauvin and D.E. Rumelhart(eds.) Backpropagation: Theory, Architectures, and Applications, Lawrence Erlbaum Associate, Publishers, New Jersey, pp.433-484(1995).
4. Rob Robertson: Examine Your Network with Ping and Traceroute(1997), http://webmonkey.wired.com/webmonkey/geektalk/97/42/index3a.html

Health Support Intelligent System for Diabetic Patient by Mobile Phone

Takumi Ichimura[1], Machi Suka[2], Akihiro Sugihara[3], and Kazunari Harada[3]

[1] Faculty of Information Sciences, Hiroshima City University,
3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194, Japan
ichimura@its.hiroshima-cu.ac.jp

[2] Department of Preventive Medicine, St.Marianna Univeristy School of Medicine,
2-16-1, Sugao, Miyamae-ku, Kawasaki 216-8511, Japan
suka@marianna-u.ac.jp

[3] Department of Mobile Multimedia, NTT DoCoMo Chugoku,Inc.,
4-1-8 Oote-machim, Naka-ku, Hiroshima-shi, Hiroshima-ken, 730-8566 , Japan
{sugiharaa,haradak}@docomo-chugoku.co.jp

Abstract. We developed the Health Support Intelligent System for Diabetic Patients (HSISD) as a web-based system. Nowadays mobile phones are widely used in our daily life and become the most popular personal communication tool. To make HSID more familiar, we add the application software for operating in mobile phone and enable data entry and patient authentification by mobile phone. In this paper, we describe the questionnaire system and the authentication system based on mobile phone technology.

1 Introduction

The Health Support Intelligent System for Diabetic Patients (HSISD) is a computer-assisted system used in the clinical field[1]. HSISD has two phases of (1) guideline-based decision support (GDS) and (2) tele-consultation (TC) as shown in Fig.1. The use of computer-assisted system like HSISD is known to be of help in making decision on treatment and in exchanging communications between a doctor and a patient. Nowadays mobile phones are widely used in our daily life and become the most popular personal communication tool. Mobile phones are not only used for calling but also for the Internet communication such as browser, e-mail, and application software. Moreover, the latest mobile phones, which are equipped with an integrated circuit(IC) card, can carry out a function of electric money. We developed HSISD as a web-based system, which is constructed with Java Servlet in the Secure Linux Server[2]. To make HSISD more familiar, we add the application software for operating in mobile phone and enable data entry and patient authentication by mobile phone. In this paper, we describe the questionnaire system and the authentication system based on mobile phone technology.

2 Questionnaires in HSISD

HSISD consists of three components as shown in Fig.2. Each clinic has its own server connected to the Internet. A patient who wants to see a doctor about his/her diabetes visits the HSISD website, registers personal information (name, address, e-mail, etc.) in

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1260–1265, 2005.

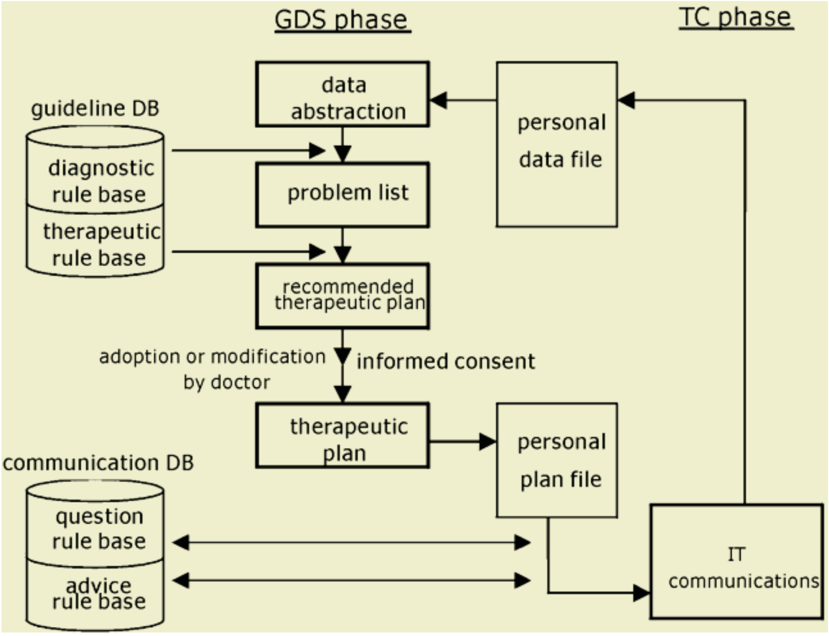


Fig. 1. Overview of HSISD

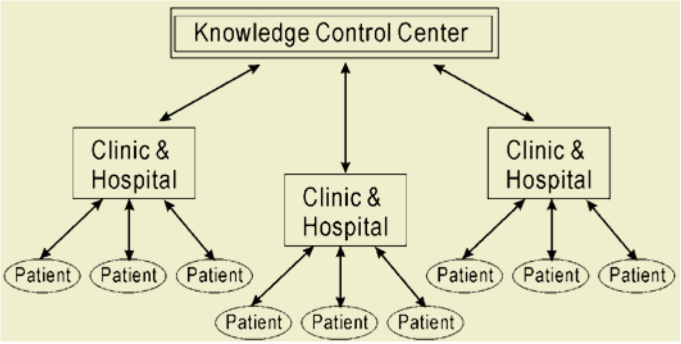


Fig. 2. Structure of HSISD

the web sheet, and selects a clinic. The selected clinic is notified of the patient through the Internet. The patient visits the specified website of the clinic and completes questionnaires about lifestyle in advance. GDS and TC are performed in each clinic.

The clinic stores the personal data in the personal data file as shown in Fig.1. GDS and TC are performed based on the personal data. The personal data file includes questionnaire data (sex, age, medical history, lifestyle, health belief, etc) and laboratory data (body mass index, blood glucose, blood pressure, triglyceride, cholesterol, urinalysis, electrocardiogram, etc). The laboratory data are input by medical staffs after medical examination. Therefore, we focus on data entry of the questionnaire data.

The patient needs to complete questionnaires about lifestyle before medical examination. However, HSISD enables data entry at home through the Internet. The web sheet for the questionnaire data is separated into 9 parts as shown in Fig.3. When the patient clicks on a part, the corresponding web form sheet is opened. Fig.4. shows the web form sheets for data entry of the questionnaire data. The patient answers the questionnaires in the web form sheets and sends the data to the clinic server. The clinic server registers the data in the database. If the patient wants to modify his/her own data, he/she accesses modification pages.

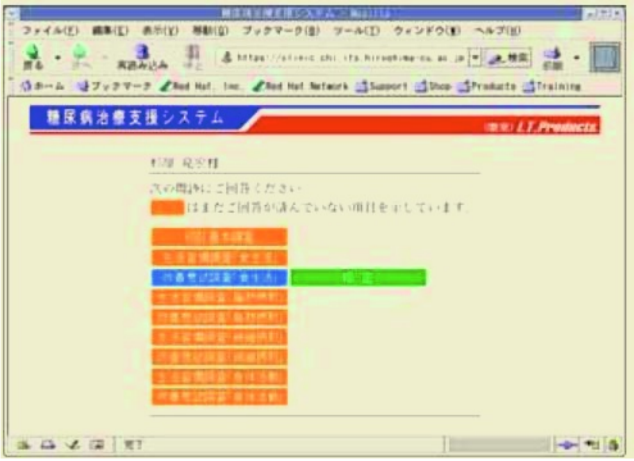


Fig. 3. Menu for data entry of the questionnaire data

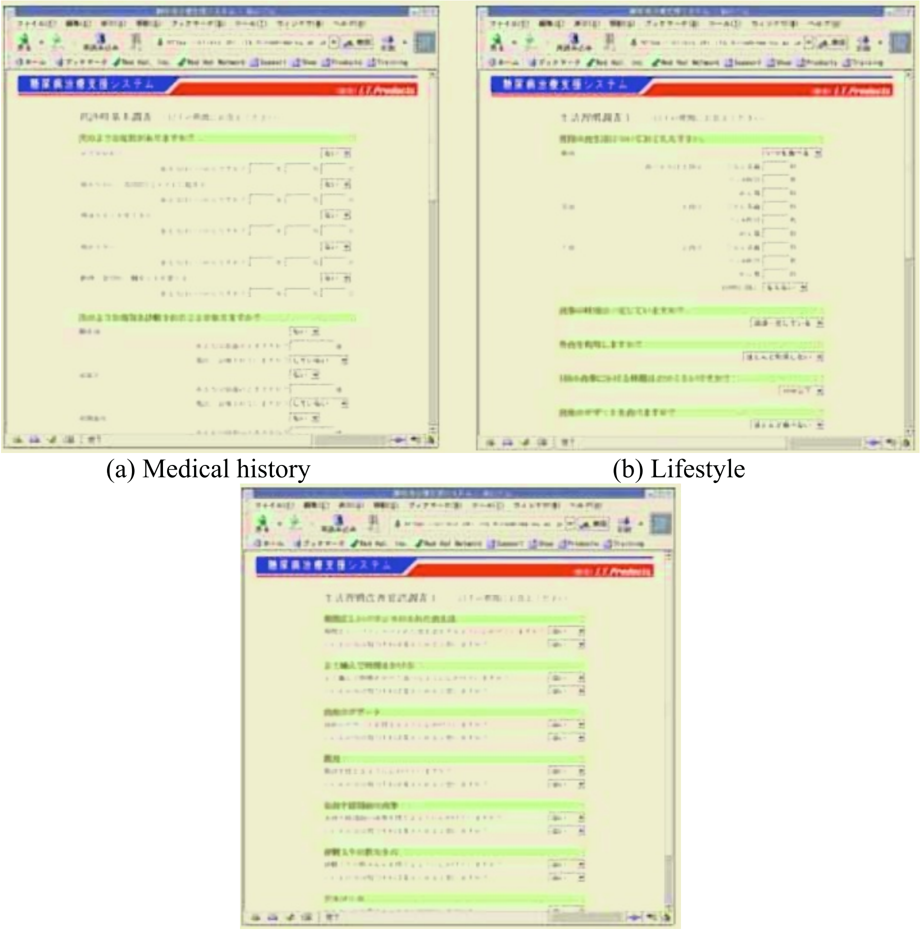
To enable data entry by mobile phone, we developed the application software for operating in mobile phone, called "i-appli[3]", produced by the NTT DoCoMo in Japan. The NTT DoCoMo has the special own network. The data packets in the network are exchanged in the perfectly secure form. Moreover, the "i-appli" has a strict limitation for sending data with TCP/IP protocol. When the patient downloads the application software from the clinic server, he/she can send data only to the clinic server having the same IP address.

The application software of HSISD, mobile HSISD, is not continuously connected to the Internet. Unless the patient answers the questionnaires and confirms to send the data to the clinic server, the mobile HSISD don't start the process of data transmission.

Fig.5 shows the displays of mobile HSISD; (a) Start menu, (b) Security confirmation, (c) Questionnaire sheet menu, and (d) Confirmation of sending data.

3 Patient Authentication by Mobile Phone with IC Card

The latest mobile phones, which are equipped with an IC card, can be carried out only a function of the Internet communication but also that of electric money. The Contactless IC card Technology, called "FeliCa[4]", developed by the Sony Cooperation is embedded into some types of mobile phones. The mobile phone with an IC card can be used for patient authentication.



(a) Medical history

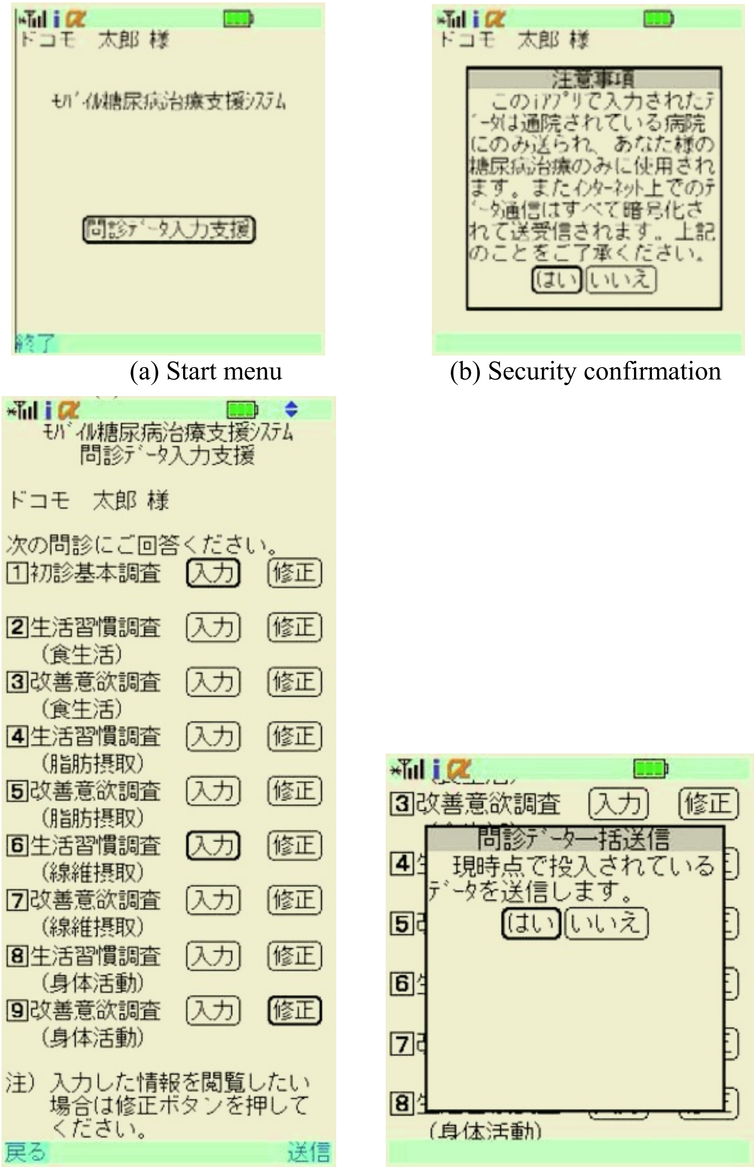
(b) Lifestyle

(c) Health belief

Fig. 4. Web form sheets for data entry of the questionnaire data

Fig. 6 shows the flow of patient authentication by mobile phone. The patient downloads the “i-appli” with the clinic ID including information of time stamp from the clinic server. The clinic ID is unique, because it is produced at each download. The encrypted data is created in the form of “<a patient’s e-mail address@the clinic ID>Password”. The password is what the patient registered as personal information.

The encrypted data is written to the IC card in mobile phone by using Software Development Kit for “FeliCa”[5] as shown in Fig. 7. The IC card consists of 2 areas; Common Area for electric money and Free Area where the user can freely use to save/load the “i-appli”. When the user brings his/her mobile phone to the clinic, the authentication is executed easily and smoothly.



ドコモ　太郎 様

ﾓﾊﾞｲﾙ糖尿病治療支援ｼｽﾃﾑ

問診ﾃﾞｰﾀ入力支援

終了

(a) Start menu

ドコモ　太郎 様

注意事項

このiｱﾌﾟﾘで入力されたﾃﾞｰﾀは通院されている病院にのみ送られ、あなた様の糖尿病治療のみに使用されます。またｲﾝﾀｰﾈｯﾄ上でのﾃﾞｰﾀ通信はすべて暗号化されて送受信されます。上記のことをご了承ください。

はい　いいえ

(b) Security confirmation

ﾓﾊﾞｲﾙ糖尿病治療支援ｼｽﾃﾑ
問診ﾃﾞｰﾀ入力支援

ドコモ　太郎 様

次の問診にご回答ください。

1 初診基本調査　入力　修正

2 生活習慣調査（食生活）　入力　修正

3 改善意欲調査（食生活）　入力　修正

4 生活習慣調査（脂肪摂取）　入力　修正

5 改善意欲調査（脂肪摂取）　入力　修正

6 生活習慣調査（線維摂取）　入力　修正

7 改善意欲調査（線維摂取）　入力　修正

8 生活習慣調査（身体活動）　入力　修正

9 改善意欲調査（身体活動）　入力　修正

注）入力した情報を閲覧したい場合は修正ボタンを押してください。

戻る　送信

3 改善意欲調査　入力　修正

問診ﾃﾞｰﾀ一括送信

現時点で投入されているﾃﾞｰﾀを送信します。

はい　いいえ

4
5
6
7
8

（身体活動）

(c) Questionnaire sheet menu　(d) Confirmation of sending data

Fig. 5. Displays of mobile HSISD

4 Conclusive Discussion

In this paper, we describe the mobile HSISD. The web-based HSISD is easily used in the environment where a personal computer is available for a long time. On the other hand, the mobile HSISD enables data entry anytime and anywhere without a personal computer. Moreover, it is likely that the user of mobile phone with IC card takes re-

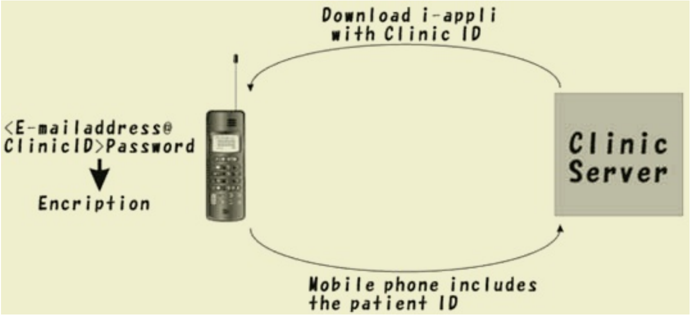


Fig. 6. Flow of patient authentication by mobile phone

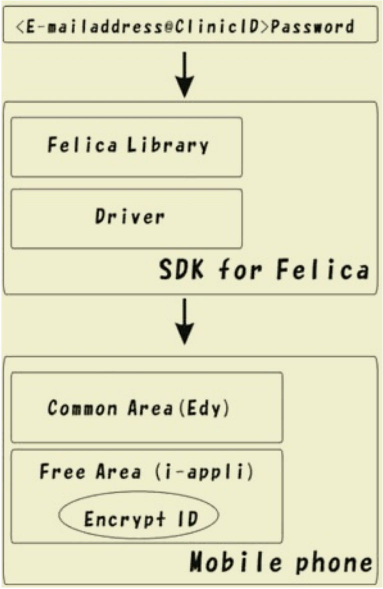


Fig. 7. Encrypted ID written to the IC card in mobile phone

sponsibility for keeping data, because the mobile phone combines a function of electric money. The authentication by mobile phone will be widely used in our daily life.

References

1. M.Suka, T.Ichimura, et al, "Health support intelligent system for Diabetic patients (HSISD)", Proc. of the 6th International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies (KES2002), Vol.1, pp.698-702(2002)
2. T.Ichimura et al., “Technologies in web based application”, Serendip Publishers, Tokyo, Japan(2004)(In Japanese)
3. i-mode and i-appli, http://www.nttdocomo.co.jp/p_s/imode/ (2005)
4. I-mode FeliCa, http://www.nttdocomo.co.jp/p_s/service/felica/(2005)
5. FeliCa, http://www.sony.net/Products/felica/(2005)

Proposal of Food Intake Measuring System in Medical Use and Its Discussion of Practical Capability

Yoshihiro Saeki and Fumiaki Takeda

Kochi University of Technology, Miyanokuchi, Tosayamada-Kochi, 782-8502, Japan

Abstract. In this paper, a food intake measuring system for medical applications is proposed. The system measures the differences of food images between pre-eaten and post-eaten, and accurately calculates the intake of calorie and nutrition. It can be an assistant of dietitians. The whole operation procedures and each component are introduced. The verification experiments of the system performance are also executed.

1 Introduction

We have been developing a measuring system of food intake for medical application [1][2]. The purpose of the development of this system is to automatize the welfare work, which includes measuring various remnant foods of patients and managing the information of them manually. It is too complicated and weary for operators. However, the information of calorie intake of patients is important for doctors because they decide the medicine intake on the base of this information.

In this paper, a measuring system of food intake using image process is proposed. The experiments of deciding the procedure of operations and evaluations are carried out[3][4]. After that, the relationship between illumination and detection of food edges is discussed. We redesigned the photographic frame for minimizing the system[5][6], and mounted wide-angle lens[7][8]. Finally, we discussed the practicability of the proposed system[9][10][11].

The management of medical meal is troublesome work, which needs dietitians to calculate the calorie intake according to the standard of food elements by comparing the intake quantities between after and before eating. The measuring results are influenceable by the individual differences of dietitians, and the efficiency of the manual operation is low. Moreover, it is difficult to realize the electronic management for patient data. There are about 80% dietitian who want a automatic instrument to reduce this kind of burden in a questionnaire survey[1].

2 The Measuring System of Food Intake

The system compares the image of foods before and after begin eaten and calculates the calorie intake of patients. First, the whole tray with foods is input

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1266–1273, 2005.

the apparatus and is taken a photo. After that, the images of dishes and foods are extracted sequentially[12][13]. Finally, the pixels of food images before and after begin eaten are compared, and the calorie intake is calculated on the base of the database of nutrition[2][3]. In this paper, the performances of each subsystem such as the dishes extraction and foods extraction are verified in emulation environments of medical center. If the proposed system is applied, it is possible to realize some functions as following, deciding standards of the measurement, decreasing the time of the measurement, managing the nutrition for patients accurately, decreasing the burden of dietitians, managing and sharing the data and measuring food images by personal digital assistance (PDA).

2.1 Structure of Hardware and Software

The appearance of the system is shown in Fig. 1. It uses 4 incandescent lamps (25W) and a camera (Resolution: 320 × 240) with a USB (Universal Serial Bus) interface begin used to communicate to Personal Computer (PC) (Specification: PentiumM 1.60GHz, 256MB RAM, WindowsXP).

The software construction of the system is shown in Fig. 2. The software of the system consists of two parts. One is image processing program, and the other is data base program (DBP). All instructions and results are transported using UDP (User Datagram Protocol). The image processing program includes the communication program, photography processing program and measuring processing program. The DBP is composed of the dish database, food menu database, food database, foodstuff database and personal database. All the parts of the program are stored in PC. The procedure including all processing from entering the menu to calculate the calorie intake is executed online.

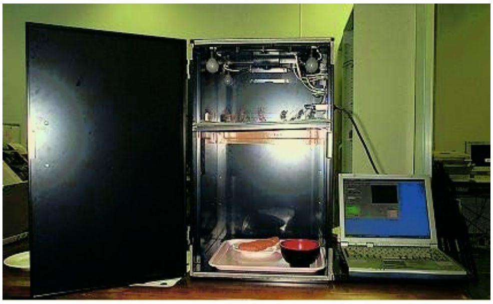


Fig. 1. System appearance

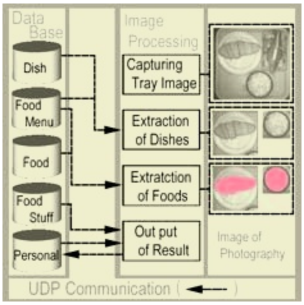


Fig. 2. System Configration

2.2 Image Processing Program and Communicating with DBP

Image processing program including communication to DBP, photography processing program and measurement processing program.

First, the communication port are initialized, and the image processing program sends commands to the DBP, and receives the information of menus and templates. Then the results of calculation are sent to the DBP.

Fig. 3 shows the image of photography processing program, in which the left image is a real-time image and the right one is the final saved image for identification. In this procedure, the image of dishes and foods are collected and calibrated (including contrast, sharpness, lightness, etc).

In the measurement processing program, the time and ID of patients are selected firstly, and then the system begin to measure the input data after the "extraction button" on the interface is pushed. If the input time and ID are not registered in the database, there will be an error message. On the contrary, the captured image is shown in the interface illustrated in Fig. 4. After that, for the captured image the transformation of corresponding image, the extraction of dishes and foods are executed by Matrox Imaging Library (MIL), witch is all-purpose for image processing, to extract dishes and foods image from captured image, and the results are output. It takes about 1 sec for the whole procedure. The surplus percentages of each kind of food and corresponding value of calorie intake, and the overall surplus percentages of foods and corresponding value of calorie intake are shown in the interface. If the food in the interface is selected, the extraction parts of foods will be shown. Therefore, it can be verified whether the extraction procedure is correct or not.

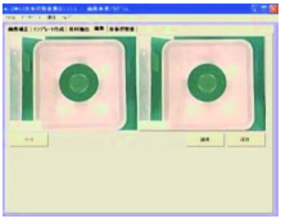


Fig. 3. Photograpy Processing

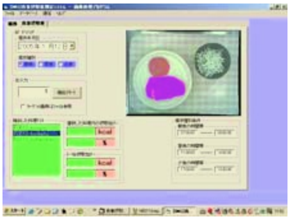
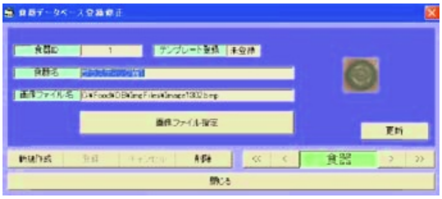


Fig. 4. Measurement Processing



食器データベース登録修正

食器ID 1 テンプレート登録 未登録

食器名

画像ファイル名 D¥Food¥DB¥ImgFiles¥Image1302.bmp

Fig. 5. Dish Database

メニューID 1

メニュー名

Fig. 6. Food Menu Database

It is essential for the system to register some kinds of data in the DBP. The measurement of captured images and calculation of the intake of calorie and different nutriments are carried out. The respective interfaces of 5 databases mentioned above are shown from Fig 5 to Fig 9.

In the dish database, the types and corresponding images of dishes, and the essential templates are included. The dish is decided on the base of this database and registered templates during the procedure of dish extraction.

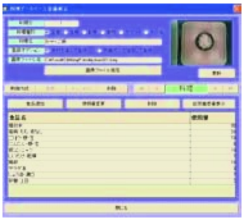


Fig. 7. Food Database

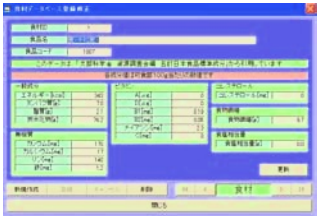


Fig. 8. Foodstuff Database

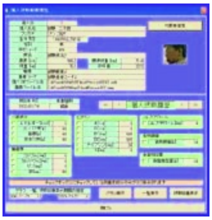


Fig. 9. Personal Database

The meal menus are registered in the menu database. After deciding the data and the name of menu, the foods, the dishes of corresponding databases. After that, it is possible to decide the differences between the measured foods and the registered foods. Therefore, it is realizable for each patient to be supplied appropriately.

First, the name and image of foods are registered in the food database, which is similar with a recipe. Then the information of the corresponding foodstuffs and seasoning are inputted from the foodstuff database, in which the data come from the standard of food composition in Japan edited by the resource investigation community of Japan.

In the personal database, the name, sex, birth, etc. of patients are registered. During the measuring procedure, the registered ID is selected from this database and the results are recorded in the corresponding database according the ID number. Furthermore, it is realizable to grasp the information of nutrition intake and historic records for individuals.

3 Result and Investigation

To improve the performance of the system, the initial light source, incandescent lamps is replaced by fluorescent lamps. The corresponding results are shown in Table 1. There are two reasons why the performance of the system is improved. One is that the amout of fluoresent light is less influenced on heat than that of incandescent light. The other is that the readiness of amout of fluorescent light higher than that of incandescent light. After that, to make the system movable, the height of the system is changed from 650mm to 390mm. The prototype and improver are shown in Fig. 10. The frame of capturing is therefore decreased correspondingly and it is difficult take a picture for the whole foods. So a wide-angle lens is replaced and it is shown in Fig. 11. The results are shown in table 2. Fig. 12 illustrates the success and failure examples of food extraction.

Table 1. Result of Light Source Experiment

Light source	Dish extraction(%)	Food extraction(%)
Incandescent	100.0	23.6
Fluorescent	100.0	80.1

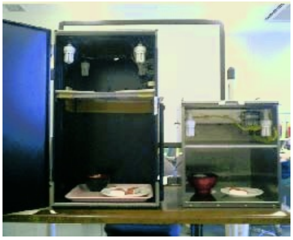
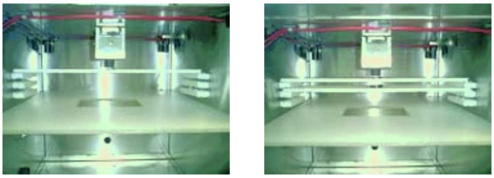
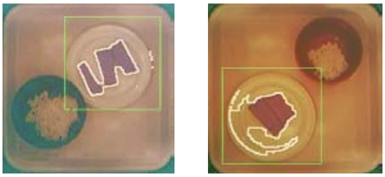


Fig. 10. Photography machine (Left:Prototype , Right:Improved)



(a)One lens (b)Two lenses

Fig. 11. Installation of wide-angle lens



(a)Success (b)Failure

Fig. 12. Measurement Experiment

Table 2. Result of Measurement Experiment

Food remainder(%)	Dish extraction(%)	Food extraction(%)
0.0	42.9	10.7
50.0	32.1	25.0
100.0	30.4	23.2
Average	35.1	19.6

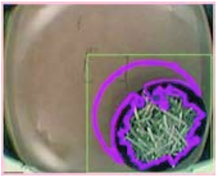
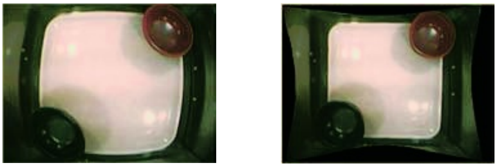


Fig. 13. Distortion of image

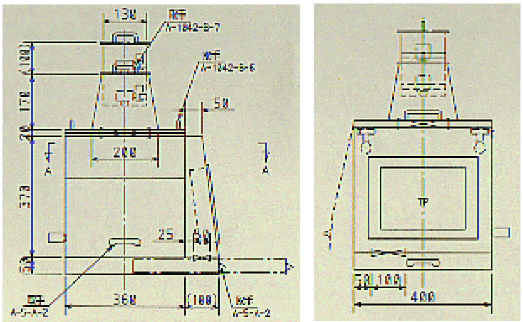


(a)before (b)after

Fig. 14. Distortion Correction Algorithm

As can be seen from table 2, the extraction capabilities are not satisfied because it occurs image distortion shown in Fig. 13 when using the wide-angle lens. To resolve this problem, the distortion correction algorithm is applied. The contrastive images are shown in Fig. 14.

As mentioned above, the size of the platform has been decreased and the corresponding problem is solved using the wide-angle lens. Another method using a telescopic structure is also applied on the system to regulate focus of the lens. The blueprint and the photo of this new platform are shown in Fig. 15 and 16. Table 3 shows the results of the contrastive experiments.



(a)Front View (b)Right Side View

Fig. 15. Blueprint of Latest Platform

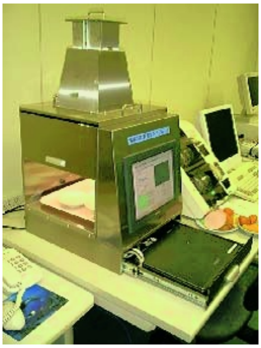


Fig. 16. System Overview

Table 3. Experimental result of extraction measurement

Platform	Dish Extraction(%)	Food Extraction(%)
Last	35.1	19.6
New	100.0	95.2

3.1 Evaluation of the System

During this section, a menu sample show in Fig. 17 is used to verify the extraction capabilities of the proposed system with different condition. The results are shown in Table 4, 5 and 6, in which the letters A, B, C, D, E and F represent pork cutlet, tomato slice, beef, julienne cabbage, rice and croquette respectively.

Table 4 reveals that in the initial 3 minutes after power-on, the system is instable. It is also revealed that the rotation angle of trays and dishes also influence the instability of extraction capabilities in Table 5 and 6. It stems from the shadow change of foods as rotation. Therefore, the effective illuminants and photographic devices are considered in future.

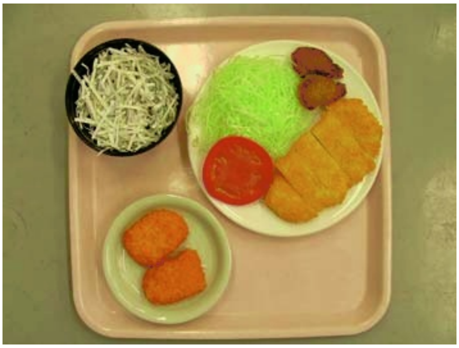


Fig. 17. Photograph of Menu

Table 4. Results of different time after power-on

Time(min)	A	B	C	D	E	F
1	×	▲	▲	▲	●	△
2	●	▲	▲	▲	●	△
3	○	▲	●	▲	●	△
4	○	▲	●	▲	●	△
5	○	▲	●	▲	●	△
10	○	▲	●	▲	●	△
15	○	▲	●	▲	●	△

Table 5. Results of different rotation angle of tray

angle(deg)	A	B	C	D	E	F
0	○	▲	●	▲	○	△
90	×	×	●	▲	○	△
180	×	×	●	▲	○	△
270	○	▲	●	▲	○	△

Table 6. Results of different rotation angle of dish

angle(deg)	A	B	C	D	E	F
0	○	▲	●	▲	○	△
45	○	●	●	▲	○	△
90	○	●	▲	▲	○	△
135	▲	▲	●	▲	△	△
180	△	▲	●	▲	△	△
225	▲	△	●	▲	△	△
270	▲	▲	●	▲	●	▲
315	×	×	○	▲	●	○

○:very good, ●:good, ▲:bad,
△:failure (dish included) and ×: failure (food included)

4 Conclusion

In this paper, a food intake measuring system was proposed and the structure of this system and the operation procedures ware also introduced. The performance of the system was improved by changed the illuminant. The evaluation experiments of the two schemes for resolve the miniaturization of the system ware executed. Then the influences of illuminant and photography parts for system performance ware discussed. In future, to further improve the performance of the system, the parts of illuminant, photography and so on will continue to be considered. We plan to use two cameras to capture 3D images of foods.

References

1. F.Takeda, H.Uchida and S.Hattori "Construction of Measuring System for Food Intake System", SPATJ, pp.75-80, 2001.
2. H.Uchida and F.Takeda "Development of a Measurement System for Food Intake using Neural Network", ICMIT'01, pp.19-24, 2001.
3. H.Kumada, F.Takeda, M.Takara "Dish Extraction Method with Neural Network for Food Intake Measuring System on Medical Use", Society of Signal Processing Applications and Technology of Japan, p21-22(2003)
4. H.Kumada, F.Takeda, M.Takara "Development and Research of Food Intake Measuring System", JACC(2003)
5. M.Takara, H.Kumada, F.Takeda "Development of Food Intake Measuring System on Medical Use", The Institute of Systems, Control and Information Engineers, p581 582, Kyoto, 2003.
6. F.Takeda, M.Takara, H.Kumada "Development of Food Intake Measuring System with Image for Medical Use", The Society of Instrument and Control Engineers, Okayama, 2003.
7. F.Takeda, "Dish Extraction Method with Neural Network for Food Intake Measuring System on Medical Use", CIMSA'03IEEE, Switzerland, 2003.
8. Y.Saeki, F.Takeda "Extension to Practical System Model on Food Intake Measuring System", ISCIE'04, p391-392

9. Y.Saeki, F.Takeda "Development of Food Intake Measuring System and Extension to Practical Model", EISS'04, p532-534
10. Y.Saeki, F.Takeda "Proposal and Extension to Practical System Model on Food Intake Measuring System", JSME'04, p205-208
11. Y.Saeki, F.Takeda "Development of Food Intake Measuring System for Medical Use and Its System Evaluation", FAN'04, p149-152
12. N.Funakubo "Pattern Recognition", Kyoritsu Publishing, pp.154-157, 1993.
13. T.Agui and T.Nagao "A Guide to Image Processing by C Language", Syokodo Publishing, pp.47-74, 2000.

Simple Web Mail System That Makes the Best Use of the Senior Citizens Social Experience

Kiichirou Sasaki[1], Yurie Iribe[2], Masato Goto[1], Mamoru Endo[3],
Takami Yasuda[1], and Shigeki Yoko[1]

[1] Graduate School of information science, Nagoya University,
{kii,masato,yasuda,yokoi}@mdg.human.nagoya-u.ac.jp
http://www.mdg.human.nagoya-u.ac.jp
[2] Information and Media Center Toyohashi University of Technology
iribe@imc.tut.ac.jp
[3] School of Computer and Cognitive Sciences, Chukyo University
endo@om.sccs.chukyo-u.ac.jp

Abstract. This research paper aims to develop a simple Web mail system for senior citizens to make the best use of their empirical knowledge as social properties, which is one plan of an informationization promotion project by industrial-government-academic cooperation begun in 2004. A simple Web mail system is made for trial purposes as a concrete example for senior citizens based on their needs, and an indicator of grappling with the current state in development and future problems is described.

1 Introduction

Because knowledge and wisdom experienced by senior citizens are accumulates they can be regarded as valuable intellectual property. Senior citizens with abundant experience and wisdom become persons with intellectual property. Therefore, in the world, Japan will become a society where life expectancy is superb; it will be rich, intellectual nation of resources [1]. However, the chance to use such intellectual resource s with Information Instruments is few due to the problem of digital divide. This digital divide problem is important by age because the ratio using Information Instruments in the generation of senior citizens decreases rapidly by Internet availability (Fig. 1) in Japan according to generation. The inside and the industrial-government-academic complex of such a situation has become integral (Fig. 2), and strategies to solve digital divide were examined. The software senior citizens can easily operate is developed. The spread activity is done as the maintenance of text and curriculum. The environment in which senior citizens wants to use Information Instruments is straightened. Use and the maintenance of the recycling personal computer as the equipment for the introduction are done. It acts as Nagoya City informationization promotion project them. In this research, a design size was arranged for the activation aim of intellectual resources by using e-mail. It brought operation sense for actually writing letters, based on senior citizens usability, and buttons and characters were considered, and a simple Web mail system was developed. As a result, because past Web mail system targeted persons generally able to handle computers,

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1274–1280, 2005.

the problem was not treated easily to be solved for senior citizens. Moreover, it was limited to changing the size and arrangements of buttons and characters, solving the problem without any mechanisms using an intellectual resources became possible in the mail system for senior citizens.

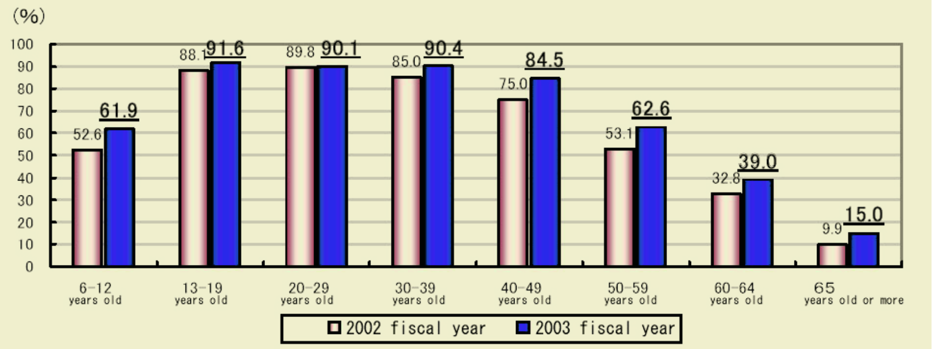


Fig. 1. Internet availability by generation in Japan. This was quoted from the communication use trend investigation in of Japan Ministry from Internal Affairs and Communication

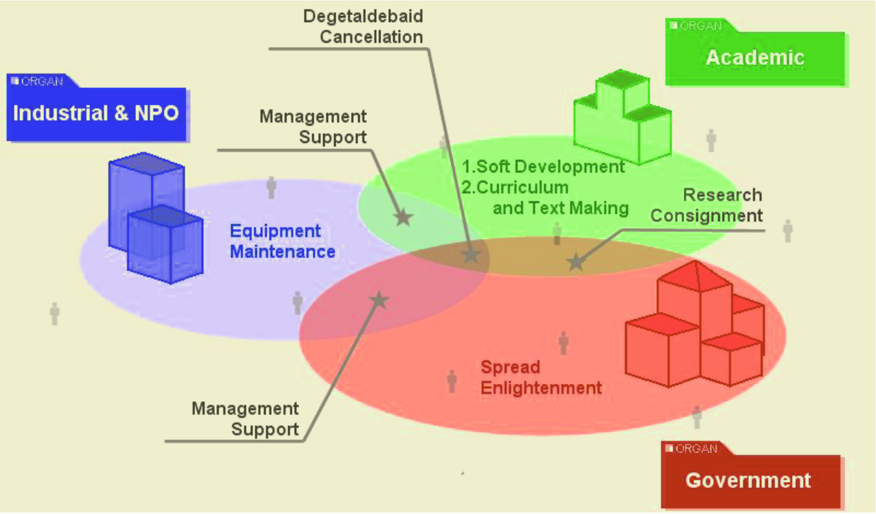


Fig. 2. Plan of informationization promotion project by industrial-government-academic complex

2 Approach of Raku Raku Pasokon e-namokun

This chapter investigates whether e-mail software is too difficult for senior citizens. A concrete settlement plan is presented. The reason for the assumption of difficulty for senior citizens was verified in a second experiment. In the first experiment it went to skill practices where it went to 2004/8/9 at the life Atsuta Ward study center and 2004/12/20 at the life Atsuta Ward study center and 2005/1/26 at the life Atsuta Ward study center. A lot of functions exist on one page, and I cannot not understood which functions I may use. Unnecessary functions that beginners don't use are provided, and so incorrect operations might be done. Functions were clarified for such problems by composing functions of minimum requirements. What, where you should inputs according to what procedures when the input work on one page increases was not understood. On the other hand, input procedure was guided by one step, and it was simplified. Moreover, the character input, problem is solved using a software keyboard separately researched and developed. E-mail addresses are not understood easily from the enumeration of the alphabet by senior citizens complicates using the addresses. On the other hand, my name and mail address can be related and used. A lot of technical terms are includes and it is difficult to understand the displayed content. On the other hand, technical terms are substituted for comprehensible terms, and the mode of expression is simplified. Because the client must upgrade the e-mail and security software, management of e-mail software is difficult. On the other hand, all functions can be used with the Web base in the direction where upgrades in the client base are lost. A system limited to functions necessary for sending and receiving basic e-mail was developed as a concrete solution. Moreover, illustrated development is done in the entire system (Fig. 3).

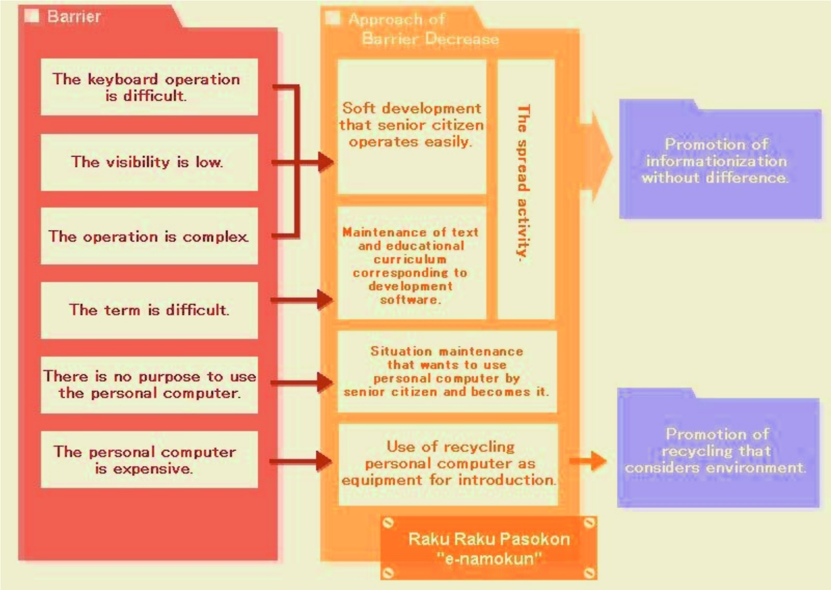

Barrier
The keyboard operation is difficult.
The visibility is low.
The operation is complex.
The term is difficult.
There is no purpose to use the personal computer.
The personal computer is expensive.
Approach of Barrier Decrease
Soft development that senior citizen operates easily.
Maintenance of text and educational curriculum corresponding to development software.
The spread activity.
Situation maintenance that wants to use personal computer by senior citizen and becomes it.
Use of recycling personal computer as equipment for introduction.
Promotion of informationization without difference.
Promotion of recycling that considers environment.
Raku Raku Pasokon "e-namokun"

Fig. 3. Approach plan of Raku Raku Pasokon e-namokun project

3 Actual Web Mail System

This section explains the four function and the achievement systems of a simple Web mail system that uses a letter model technique based on mailer usability for senior citizens. E-mail is made easy to use with the letter model technique here by losing the disgust to the computer by bringing it close to the communication of the letter.

3.1 Web Function to Achieve Letter Model Technique

When the letter model technique is achieved, it is necessary to assume the e-mail system of the Web base. This has a system upgrade and a problem of dependence etc, by the client machine in the mailer of past client base. Therefore, a lot of favors that solve by e-mail the Web base and depend on it are obtained. Mail is sent and received with Web, and it will be necessary to anticipate future changes in the connected database for scale expansion and plug in. Therefore, PHP was adopted from high-speed operations compared with quality and other server side scripts of abundant sheath database convertibility of the kinds of in a library [2]. Moreover, PHP can facilitate the shift to the adoption of Squirrel Mail. Which is a typical Web mail system. The Windows platform has a comparative advantage. As a result, it is thought that NPO and the volunteer group can easily do function enhancing and the management in the future. The database adopted MySQL to forecast the scale of the experiment course in the future. The load on the user side has been reduced by unitary managing user information and the address book by the database. As for the data shift between the communication of data by Web applications and page, the users distinction is necessary. This distinguished users by the session function of PHP and handed over necessary data.

3.2 Securing of Portability by Web Mail

The method of receiving e-mail from the mail server to the mailer on a terminal as a method of sending and receiving individual mail using POP3 is a main topic. However, ease of opportunity and introductions that don't use an individual terminal this time but use a public terminal was included in the specifications. Therefore, portability was achieved by constructing Web mail that transmitted e-mail by using IMAP, receiving e-mail, and SMTP (Fig. 4). Moreover, it is anticipated that using a Web mail system such as Squirrel Mail will receive attention, and the e-mail environment by web mail will become a subject in the future. E-mail can be easily managed, and in this background, the environment is enumerated and maintained for which an Internet browser can be used securely. Therefore, it became a system into which many kinds of personal computers could be introduced in the computer introduction promotion that used recycled computers of the e-namokun project.

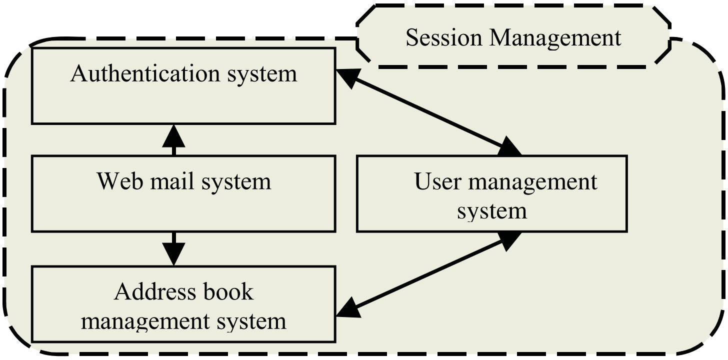
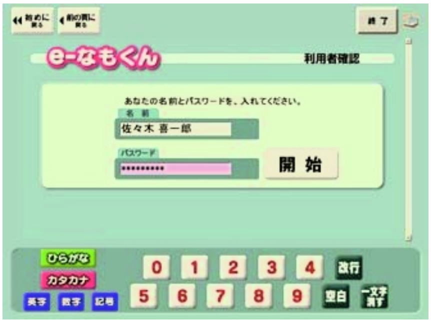


Fig. 4. e-namokun mail system function chart

3.3 Securing of Portability by Web Mail

The pop up screen of attestation made with an attestation mechanism of a browser cannot change the design because it depends on a browser. Therefore, it is necessary to make an original attestation mechanism (Fig. 5) and consider the design of the attestation screen and operation. This system achieved it by constructing an authentication system that used the session function of PHP. The attack measures misuse the cross-site script, JavaScript, and VBScript. It achieved the data script of user input by excluding the tag of the character string and a special character according to the htmlspecialchars() function. Data falsification measures used cross-check and the PHP session management function of the sent data. Moreover, using SSL for communicate prevents the appropriation measures of session ID. Random attack measures of session ID are scheduled to be given. Because mail is sent and received, the computer virus problem of e-mail communication tends to be strong in Trojan horse type worms in the attached files and the script type virus of the Web base. Moreover, measures are include to manage the server side for computer viruses [3]. Therefore, the e-namokun mail system achieved mail sending and receiving functions and the address book with a Web base (Fig. 6). As a result, the cost of measures to my computers can be reduced for senior citizens [4][5].

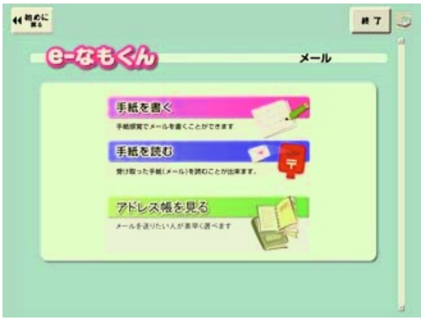


あなたの名前とパスワードを、入れてください。

名前: 佐々木 喜一郎

パスワード: **********

Fig. 5. Attestation screen



手紙感覚でメールを書くことができます

受け取った手紙(メール)を読むことが出来ます。

メールを送りたい人が素早く選べます

Fig. 6. Function selection screen

4 Evaluation Experiment

Three people in their fifties, four in their sixties and seven in their seventies evaluated the e-namokun mail system by questionnaire. There were eight men and six women for a total of fourteen people in the first experiment. The e-namokun mail system obtained evaluations in which many people could answer the question how e-mail treated the sense of actually exchanging letters. However, one opinion didn't not spring actually feel whether a sent e-mail actually reached its destination clearly an item peculiar to e-mail such as title etc and function must be improve. It asked eight of the users to compare past e-mail software and the e-namokun mail system. Seven people answered that past electronic mail software was easier, and one person said that the e-namokun mail system was easier. When opinions were requested from about which past electronic mail software was easy to use, some people mentioned that without the enhanced feature such as address books past electronic mail software included. The sending and receiving of basic e-mail became the results when a lot of functions were requested by experts though it easily used the e-namokun mail system. However, a person who has never used a computer at all is assumed, and the target of the e-namokun mail system wants to solve this problem with additional function in the future.Three people in their fifties, four in their sixties and seven in their seventies evaluated the e-namokun mail system by questionnaire there were eight men and six women for a total of fourteen age. It brought it close to the sense of exchanging an actual letter by excluding a peculiar item of the e-mail system such as the title. Words referring to opinions were received in the first experiment and displayed the sent messages. As a result of evaluations, the e-namokun mail system concluded that a lot of people could answer the questions that e-mail was able to treat by a sense of actually exchanging the letter. It questioned eight the users to compare past electronic mail software and the e-namokun mail system for simplicity. Seven people answered that past e-mail software was user easier, and one person said that e-namokun mail was easier. It is a current state rarely used for usual communication, even though the address book and reply functions were added, on the other hand. I want to operate it as an evolving e-mail system that gives various directions by enhanced features. In the bearing of NPO and volunteering the enhanced features and management, systems that demonstrate real value only after synchronized as a whole for the e-namokun system in the form put together on each occasion.

5 Conclusions

To use the huge knowledge of senior citizens, the use of Information Instruments is indispensable. However, senior citizen is used by digital divide by age, and Information Instruments cannot use the offer, accumulation, and information collection and store-and-forward processing of knowledge. It paid attention to the e-mail of the most important knowledge use tool in this research when the knowledge society such as uniting information exchange and knowledge was composed. E-mail was offered to senior citizens in an easy shape, and the knowledge not codified of expression was sometimes absorbed in collaboration by e-mail. A mechanism to use it similarly was proposed. Moreover, I want to develop a synthetics system available in the institution of the problem based on a lot of opinions of persons using it as social knowledge, concerning education, various opinions, the arrangements of answer, and the embarrassments of other people. It is thought to have been achieved for the first time when working on details that became the developments of the e-mail system not so far from this by radical industrial-government-academic cooperation. This was achieved by substantial opinions in the site caused by NPO that supports the approach from the side of educational institutions named the function of the e-mail that doesn't arise under inter-enterprise competition, the university, the administration and the volunteers. I want to try to use social knowledge in an approach to the industrial-government-academic complex in the future.I got large cooperation from everybody at Nagoya City related to the informationization promotion project (named: Raku Raku Pasokone-namokun business) for Nagoya City senior citizens when this research was advanced, the Nagoya City Board of Education, the Nagoya Foundation Urban Industry Promotion Public Corporation, and the Promotion Mechanism of

NPO IT eco-cycle. Part of this research was funded by the Nagoya Foundation Urban Industry Promotion Public Corporation Joint Research, the Ministry of Education, Culture, Sports, Science and Technology science research expense subsidy, and the Ministry of Education, Culture, Sports, Science and Technology private university high technology research central subsidy. It depends on the COE program "Intellectual integration (IMI) of the voice images for the social information base" in the Ministry of Education, Culture, Sports, Science and Technology 21st century.

References

1. Martin, P.T.; Email and the Internet as a teaching tool: a critical perspective Frontiers in Education Conference, 1996. FIE '96. 26th Annual Conference., Proceedings of Volume 2, 6-9 Nov. 1996 Page(s):823 - 825 vol.21. Kaushik, S.; Ammann, P.; Wijesekera, D.; Winsborough, W.; Ritchey, R.; A policy driven approach to email services. Policies for Distributed Systems and Networks, 2004. POLICY 2004. Proceedings. Fifth IEEE International Workshop on 7-9 June 2004 Page(s): 169 - 178
2. Pendharkar, P.C.; Young, K.; The development of a construct for measuring an individual's perceptions of Email as a medium for electronic communication in organizations. Professional Communication, IEEE Transactions on Volume 47, Issue 2, June 2004 Page(s): 130 – 143
3. InSeon Yoo; Ultes-Nitsche, U.; How to predict e-mail viruses under uncertainty Performance, Computing, and Communications, 2004 IEEE International Conference on2004 Page(s): 675 - 679
4. Hsien-Hau Chen; Yung-Sheng Chen; Hsia-Ling Chiang; Chung-Huang Yang; Design and implementation of smart card-based secure e-mail communication. Security Technology, 2003. Proceedings. IEEE 37th Annual 2003 International Carnahan Conference on 14-16 Oct. 2003 Page(s): 225 - 231
5. Balzer, R.; Assuring the safety of opening email attachments. DARPA Information Survivability Conference & Exposition II, 2001. DISCEX '01. Proceedings Volume 2, 12-14 June 2001 Page(s): 257 - 262 vol.2

Evaluating Navigation History Comparison

Koichi Ota and Akihiro Kashihara

The University of Electro-Communications 1-5-1 Chofugaoka, Chofu,
Tokyo, 182-8585 Japan
{kou,kasihara}@ice.uec.ac.jp

Abstract. Self-directed navigation involving knowledge construction in learning Web contents, called navigational learning, is not so easy for learners since they often fail to notice the pages to be learned and semantic relationships between the pages navigated. It is accordingly necessary to enhance their awareness of the incompleteness of knowledge that they have constructed. We call such awareness unknown awareness. However, it is hard for learners to gain it by themselves. We have developed the system, which enables them to compare their navigation history to others, and which can select some navigation histories more appropriate for the history comparison from a repository. This paper describes a case study with the system whose purpose is to ascertain if the history comparison contributes to providing an unknown awareness.

1 Introduction

In a self-directed navigation process involving knowledge construction in hyperspace, learners often finish learning with incomplete knowledge. In this work, we address this issue by means of navigation history comparison [6].

Web contents generally provide learners with hyperspace where they can navigate in a self-directed way. It involves constructing knowledge, in which they would make semantic relationships among the pages learned in the navigation process [7]. In this paper, the self-directed navigation with knowledge construction is called navigational learning. On the other hand, the navigational learning is not so easy for learners since they cannot always find a navigation path to be followed for achieving their learning goal [7].During navigation, they also often fail to notice the pages to be learned and semantic relationships between the pages navigated, and their navigational learning could finish with incomplete knowledge even if they think the knowledge construction is complete [3]. In order to facilitate the navigational learning, it is necessary to enhance their awareness of navigation path that they should follow and of the incompleteness of knowledge that they have constructed. We call such awareness *unknown awareness*. The *unknown awareness* is a key to success in the navigational learning, it is hard for learners to gain it by themselves [5],[6].

Our approach to this issue is to compare their navigation histories to others to know the incompleteness of knowledge they have constructed. The important point towards providing the unknown awareness is how to represent navigation history. It is hard for learners to share/reuse navigation history including a sequence of the navigated Web pages, which is produced by Web browsers. This suggests the necessity to represent navigation history so that the knowledge construction process can be made clear.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1281–1287, 2005.

This paper discusses how to reuse navigation histories to provide learners with an *unknown awareness* in navigational learning on the Web. In order to make navigation histories reusable, we first use an interactive history, called *IH*, which we have developed for helping learners reflect on their own navigational learning processes [4]. It encourages learners to annotate their navigation history with the knowledge construction process. *IH* allows learners to make a link between any navigated pages to draw their knowledge construction process by themselves.

We have also developed *ihComparator*, which enables the learners to become aware of the incompleteness of their constructed knowledge from the difference between their own navigation history and the histories properly selected by *ihComparator*. The history comparison enables them to gain an unknown awareness in self-directed navigational learning of Web contents.

This paper describes a case study with *ihComparator* whose purpose is to ascertain if the history comparison provides an unknown awareness. The results of this study suggest that *ihComparator* can give learners unknown awareness in a proper way.

2 *ihComparator*

2.1 Framework

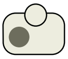
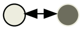
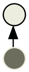
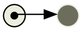
ihComparator compares navigation histories generated with *IH*. In comparing learners' navigation history to others, the large difference does not give them useful information since their navigational learning process is quite different from the others' one. It is accordingly to select a navigation history generated by others, which is analogous to the learners' history, for comparison. *ihComparator* selects a profitable history from a repository including the navigation histories in which the same contents are learnt with the same purpose. In the following, we describe the functionalities of IH and *ihComparator*.

2.2 Interactive History

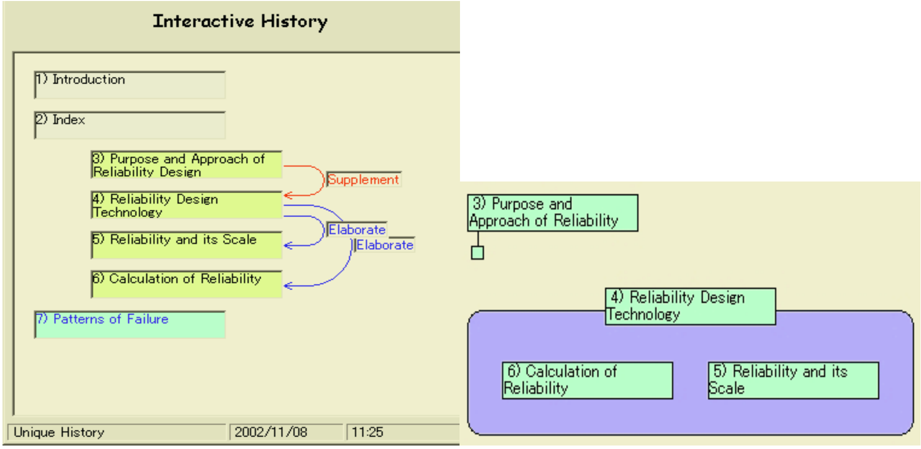
IH has two functionalities, which are annotated navigation history and knowledge map. *IH* enables learners to annotate a navigation history, which includes Web pages sequenced in order of time they have visited, with knowledge construction processes. In this annotation activity, they can make a link between any navigated pages. A navigation goal representing the semantic relationship between the pages is attached to the link. The navigation goal is classified into six as shown in Table 1. The process between the linked pages can be viewed as the one of achieving the navigation goal. We call it the navigation goal primary navigation process. Carrying out several primary navigation processes, they would construct their knowledge. The learners can look at the annotated navigation history on their demand during navigation to reflect on their navigational learning processes.

IH system next generates a knowledge map that visually represents semantic relationships among Web pages included in primary navigation processes in the annotated navigation history. *IH* first transforms each primary navigation process into a visual representation by means of the scheme as shown in Table 1. This table shows the correspondence of a navigation goal to a visual representation of the semantic relationship between the starting and terminal pages [4]. *IH* second generates a knowledge map by combining visual representation of each primary navigation process.

Table 1. Navigation Goals and Visual Representation

Navigation Goals	Visual Representation	
Supplement	Inclusion	
Elaborate	Set	
Compare	Bidirection arrow	
Justify	Vertical arrow	
Rethink	Superposition	
Apply	Arrow	

○ Starting page ● Terminal page



Interactive History

1) Introduction

2) Index

3) Purpose and Approach of Reliability Design

4) Reliability Design Technology

5) Reliability and its Scale

6) Calculation of Reliability

7) Patterns of Failure

Supplement

Elaborate

Elaborate

Unique History | 2002/11/08 | 11:25

3) Purpose and Approach of Reliability

4) Reliability Design Technology

6) Calculation of Reliability

5) Reliability and its Scale

(a) (b)

Fig. 1. An Example of Annotated Navigation History and Knowledge Map

Figure 1 shows an example of annotated navigation history and knowledge map. The knowledge map (a) is generated from Figure 1(b). The knowledge map generally consists of several islands including some primary navigation processes. We call them knowledge islands.

2.3 *ihComparator*

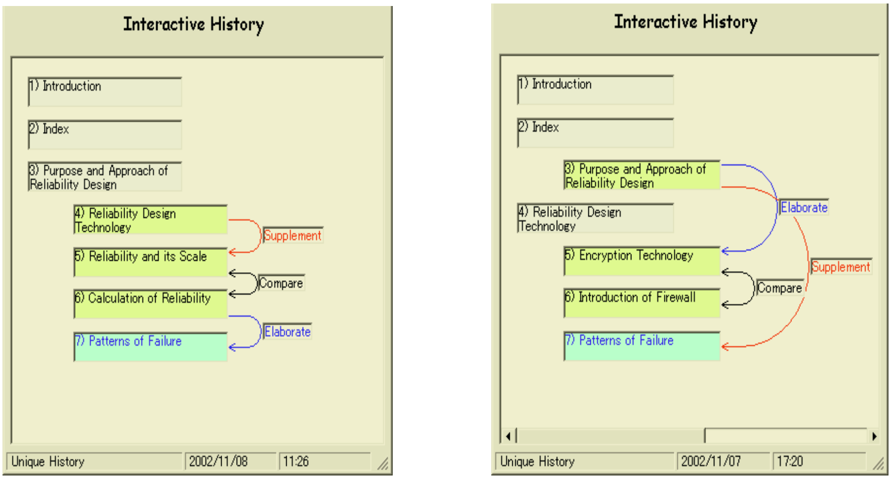
ihComparator provides two functionalities, which are history selection and visualization for history comparison. *ihComparator* first selects a history from a repository, which accumulates annotated navigation histories generated with IH. *ihComparator* then visualizes the difference between the selected history and learners' history so that they can become aware of it without difficulty.

In order to give learners an unknown awareness in a proper way, *ihComparator* selects a navigation history generated by others, which is analogous to the learners' history, for comparison. We currently expect learners to have two learning effects from the history comparison, which are knowledge deepening and knowledge widening. The knowledge deepening effect expects them to become aware of new semantic relationships between any two pages that are included in their primary navigation processes. The knowledge widening also expects them to become aware of new page to be learned that are not included in their primary navigation processes. *ihComparator* selects a history from the repository so that learners can have the knowledge deepening effect or knowledge widening effect from the difference between the selected history and their own history. In case of providing the knowledge deepening effect, *ihComparator* selects a history by fulfilling the following conditions:

a. High similarity: there exist more common pages in the primary navigation processes in both histories and there exist the same semantic relationships among the common pages, and

b. High unknown: there exist more semantic relationships between common pages in the primary navigation processes in both histories, which are unknown to learners.

For example, Figure 1(a) shows a navigation history generated by a learner. Figure 2(a) and 2(b) shows others' navigation histories generated from the same Web contents and the same learning goal. His/her navigation history has four pages, which are included in the primary navigation processes. The history shown in Figure 2(a) also has four pages, of which three entitled *Reliability Design Technology*, *Reliability and its Scale*, and *Calculation of Reliability* correspond to the pages in Figure 1(a). The history shown in Figure 2(b), on the other hand, has four pages included in the primary navigation processes, of which one entitled *Purpose and Approach of Reliabil-*



(a) (b)

Fig. 2. Navigation Histories for Comparison

ity Design corresponds to the page in Figure 1(a). In this case, the history in Figure 2(a) is accordingly more analogous to his/her history.

In case of providing the knowledge widening effect, *ihComparator* also selects a history by fulfilling the following conditions:

a. *High similarity***:** in addition to the high similarity for providing the knowledge deepening effect, there exist more common primary navigation processes in both, and

b. *High unknown***:** there exist more pages unknown to learners, which are included in the primary navigation processes.

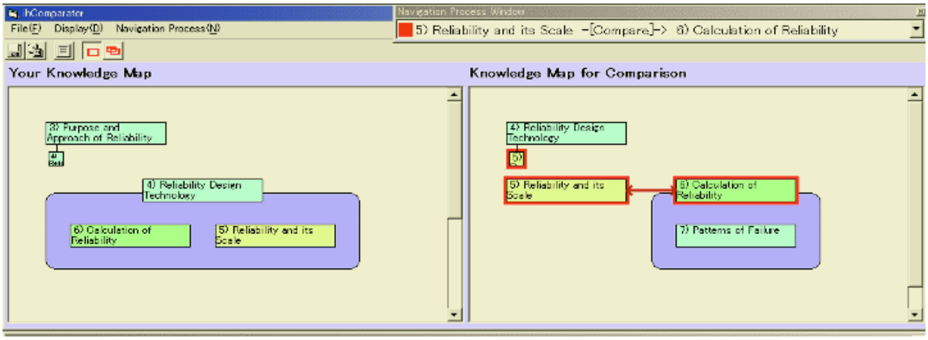

ihComparator
File(F) Display(D) Navigation Process(N)
Navigation Process Window
5) Reliability and its Scale -[Compare]-> 6) Calculation of Reliability
Your Knowledge Map
Knowledge Map for Comparison
3) Purpose and Approach of Reliability
4) Basi
4) Reliability Design Technology
6) Calculation of Reliability
5) Reliability and its Scale
4) Reliability Design Technology
5)
5) Reliability and its Scale
6) Calculation of Reliability
7) Patterns of Failure

Fig. 3. User Interface of *ihComparator*

In order to make the comparison between navigation histories easy, *ihComparator* next generates knowledge maps from learners' navigation history and the selected history, and then displays them so that the difference between the maps is highlighted.

Figure 3 shows the knowledge maps that are generated from Figure 1(a) and 2(a). Two primary navigation processes, which are unknown to the learner, are highlighted in the knowledge map generated from Figure 2(a).

3 Evaluation

In order to ascertain whether *ihComparator* selects a navigation history appropriate for providing unknown awareness, we have had a case study. In this study, we prepared a repository, which included navigation histories obtained from 16 learners who learned the same Web contents with the same learning goal by using *IH*. Subjects were 16 graduate and undergraduate students who were unfamiliar with domain knowledge (about stock investment) represented in the Web contents. The number of the pages was 98. It had a quite complex hyperspace.

The procedure of this study with each subject was as follows. The subject was first required to use *IH* to learn the Web contents with the learning goal. The annotated navigation history generated was accumulated in the repository. Two days later, the subject reviewed the knowledge map generated from his/her annotated navigation history, and the Web contents.

The subject was then given two knowledge maps called Map-with*ihComparator* and Map-without*ihComparator*, which were selected from the repository. Map-with*ihComparator* was selected by *ihComparator*, which was intended to provide him/her with the knowledge deepening/widening effect. Map-without*ihComparator* was selected by hand, which was less analogous to his/her navigation history, and which included the same number of primary navigation processes unknown to him/her as Map-with*ihComparator* included. He/she was then required to carefully compare his/her knowledge map to Map-with*ihComparator* and to Map-without*ihComparator*, and then to decide which map was fruitful for him/her to deepen/widen his/her knowledge, provided that he/she was not informed beforehand which map was selected by *ihComparator*. He/she was also required to mark the part in the selected map, in which he/she gained unknown awareness. The questionnaire was finally conducted.

As a result of the study, 13 subjects (81.25%) decided that Map-with*ihComparator* was more fruitful than Map-without*ihComparator* about the knowledge deepening effect. This suggests that the navigation history selection in *ihComparator* is appropriate for learners who try to gain the unknown awareness from a knowledge deepening point of view. From the questionnaire, there were a lot of opinions that it was easy to become aware of semantic relationships unknown since the knowledge islands in the selected map included the same pages as the knowledge islands in his/her knowledge map. Moreover, the part where the subjects marked corresponded to the one that *ihComparator* highlighted. This suggests that visualization for the history comparison works effectively.

As for the knowledge widening effect, however, 10 subjects (62.5%) decided that Map-with*ihComparator* was more fruitful than Map-without*ihComparator*. The part where they marked also corresponded to the one that *ihComparator* highlighted. On the other hand, we got an interesting opinion from all subjects who chose Map-without*ihComparator* as follows: it was easy to become aware of the pages unknown since these pages consists of a knowledge island in Map-without*ihComparator*. In fact, the pages unknown to them in Map-without*ihComparator* were almost twice the ones in Map-with*ihComparator*. But, these pages consisted of a knowledge map. That is the reason why they can gain the unknown awareness. This suggests another possibility of the knowledge widening effect that learners can widen their knowledge not only by finding new pages that could be related to their primary navigation processes, but also by finding new knowledge islands.

4 Conclusion

This paper has described a case study with *ihComparator*, which enables learners to compare their navigation history to others. The results of the study suggest that *ihComparator* can provide them with knowledge deepening and widening effects as unknown awareness in a proper way.

In future, we will refine ihComparator to help learners more properly gain unknown awareness and to promote their navigational learning process.

References

1. Brusilovsky, P. Methods and Techniques of Adaptive Hypermedia, Journal of User Modeling and User-Adapted Interaction, 6, pp.87-129 (1996).
2. Dillon, A., McKnight, C, and Richardson, J. Space-the Final Chapter or Why Physical Representations are not Semantic Intentions, in McKnight, C., Dillon, A., and Richardson, J. (eds): HYPERTEXT A Psychological Perspective, Ellis Horwood Limited, pp.169-191 (1993).
3. Hammond, N. Learning with Hypertext: Problems, Principles and Prospects, in McKnight, C., Dillon, A., and Richardson, J. (eds): HYPERTEXT A Psychological Perspective, Ellis HorwoodLimited, pp.51-69 (1993).
4. Kashihara, A., Hasegawa, S., and Toyoda, J. An Interactive History as Reflection Support in Hyperspace, Proc. of ED-MEDIA 2000, pp.467-472 (2000).
5. Kashihara, A., Hasegawa, S., and Toyoda, J. Adaptive Navigation Path Previewing for Learning on the Web, Proc. of AH2002, pp. 518-521 (2002a).
6. Kashihara, and Hasegawa, S. Unknown Awareness in Navigational Learning on the Web, Proc. of ED-MEDIA2004, pp.1829-1836 (2004).
7. Thuering, M., Hannemann, J., and Haake, J.M. Hypermedia and Cognition: Designing for Comprehension. Communication of the ACM, 38, 8, ACM Press, pp.57-66 (1995).

A Visualization System for Organizing and Sharing Research Information

Youzou Miyadera[1], Naohiro Hayashi[1],
Shoichi Nakamura[2], and Setsuo Yokoyama[1]

[1] Tokyo Gakugei University, 4-1-1, Nukui-Kita, Koganei, Tokyo 148-8501, Japan
[2] Fukushima University, Kanayagawa 1, Fukushima, 960-1296, Japan

Abstract. A trade-off exists between the management of research information based on personal viewpoints and the sharing of it on a large scale. To resolve this, we have developed methods for managing research information. First, we propose use of a transition graph to visually express research information and the relationships among the information. Second, we describe a method that enables knowledge sharing on a large scale by converting research information expressed with a transition graph into a common view with specified constraints on their redrawing. Third, we discuss a research support system that applies the proposed methods.

1 Introduction

Colleagues within an organization such as a university laboratory often tackle related research targets. Information created through the activities of others, as well as open documents, can make a valuable contribution to such related research. For instance, the body of work that other members have referred to, the part of their research they have applied the references to, and the actual progress in their research enabled by those references can be very informative. Therefore, it is important to collect and share all possible information generated in research activities. However, it is difficult to manage research information systematically as the research evolves. Likewise, it is difficult to share the information accumulated by each member on an organization scale.

Many approaches to the accumulation and sharing of such information have been reported. Trial systems to assist in the accumulation and sharing of know-how within different types of organization have also been reported [1] [2] [3] [5] [6]. There are certain strategic keys to effective management of research information within an organization. First, individuals should be able to organize the information and understand how the research is evolving based on their personal viewpoints. Second, systemized information must be effectively shared among the members of the organization. A mechanism to promote the creation of new ideas through the sharing of research progress is also important. Unfortunately, many of the approaches mentioned above focus only on information sharing on an organization scale. On the other hand, the approaches targeting the personal management of information do not consider its sharing among members of an

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1288–1295, 2005.

organization. This illustrates the trade-off we have to contend with - to manage research information based on personal viewpoints, free arrangement with the fewest possible constraints according to each researcher's needs is desirable; in contrast, an orderly structure with constraints corresponding to specified regulations is better for information sharing on an organization scale. Thus, the trade-off is between the benefits of personal management of information and those of sharing it among members of an organization according to standardized rules. For a research organization, both forms are needed for effective management of research information.

Our aim is to resolve this trade-off. To do this, we first propose a research information transition graph (RITG) that enables each researcher to arrange research information in an individual field and manage it according to the course of his or her own research. Moreover, we propose a method for converting the RITGs created by others into a common view and converting them based on the specified regulations of an organization. This method enables the sharing of research information among members within a research organization. The RITG is drawn on a two-dimensional plane. Each node of an RITG expresses research information and each edge indicates a transition between the respective nodes. Each node stores the kind of information, the author, and the date and time of creation as attributes. Each edge stores the kind of transition.

In this paper, we first explain the definition of an RITG and how research information is classified. We also propose a method for managing research information and a method for converting RITGs into a common view. In addition, we describe a support system in which these proposed methods are applied. This system enables researchers to manage research information based on their personal viewpoints while being able to review the whole body of research information in their field. The merits of this system are that it helps researchers to understand the important developments in other's research, discover new information by taking advantage of those developments, and easily share and effectively apply all available research information.

2 Current Research Activities and Problems

Our goal is to enable researchers to manage the information from research activities and new research developments based on their personal viewpoints while also promoting information sharing. As a first step, we analyzed how research activities are done at present. Research activities tend to consist mainly of particular core tasks (e.g., surveys of related research, collection of research information related to one's own research, examination of collected information and comparison with one's own research, arrangement of ideas, refinement of the direction and methods of one's research, seminar presentation and discussion, writing of academic papers) centered on each research target. Research information can take various forms, including the academic papers of others, one's own ideas, and comments by others.

At first, classifying and organizing research information can be difficult since it is scattered and exists as electronic documents, printed documents, and not yet fully developed ideas (*problem 1*). Moreover, how relationships among research information change and develop over time becomes increasingly difficult to grasp as the quantity of information grows (*problem 2*). It is also important, but difficult, to grasp how the evolution of pertinent information relates to one's own research and to manage the overall development of research (*problem 3*). Solving problems 1 to 3 leads to the management of research information based on personal viewpoints.

Within a research organization, though, the work of a researcher is often related to that of other researchers, so the information from one member's research should be shared among colleagues. However, clearly expressing how one's research has taken new directions and promoting sharing of the consequent information can be difficult (*problem 4*). Solving problem 4 leads to the sharing of research information organized on an organizational scale.

If the management of research information based on personal viewpoints has priority, it becomes difficult to share information among members of an organization. Conversely, giving priority to the sharing among members of the organization restricts the free management of research information according to individual viewpoints. Thus, there is a trade-off between the management of research information based on personal viewpoints and the sharing on an organization scale. If we can solve these four problems simultaneously, though, we can promote both personal management and sharing among members of an organization. Such a solution should make research activities much more productive.

3 Organizing Research Information with a Transition Graph

Research information should be organized based on personal viewpoints to solve problems 1 to 3. Therefore, in this research we have tried to make it easier to grasp the meaning of research information and its evolution by applying a visualization method [4]. This method enables the free arrangement of research information in individual fields by expressing the information and its transitions as, respectively, the nodes and edges of a graph.

To find out how the information created through research activities and the transitions in such information are classified, we surveyed fifteen researchers through a questionnaire and oral interviews. Each of these subjects had at least three years of experience as a researcher. Each subject was asked to list the classifications that applied to the information created through their research activities and the transitions in their research assuming that the research transitions could be fit into a two-dimensional plane. When we sorted the survey results, we obtained the typical classifications shown in Tables 1 and 2.

We propose a method of using a research information transition graph (RITG) to draw together all the information created through research activities in a personal field. Each node of the transition graph holds the classification of research information (Table 1) and each edge holds a classification of information transition (Table 2) as their respective attributes. This graph is a directed one since

Table 1. Classifications of research information

Classification	Author	Meaning
Related research	Others	Reference to related research
Product	Own/Others	Own or others' products(e.g., an academic paper)
Valuable information	Own/Others	Memorandum obtained fromspecified information
Problem	Own/Others	Problem being held byspecified information
Solution	Own/Others	Solution to specified problem
Idea	Own/Others	Own or others' idea
Detailed information	Own/Others	Detailed content of specifiedinformation
Other information	Own/Others	Other information added tospecified information

Table 2. Classifications of research information transitions

Classification	Meaning
Update&progress	Updating and expanding the information
Origin	Origin of created information
Application	Applying the information

an edge indicates the transition of information. Fig. 1 shows an example of an RITG. Each researcher has an individual field as shown and draws an RITG based on their original viewpoints.

Examples of the research transitions from Fig. 1 are: [Related research 1] holds [Problem 1], [Product 1] holds [Problem 2], [Product 2] developed into [Product 3] with [Related research 6] as a reference, [Product 3] currently holds [Problem 4], the research can be extended since no edge starts from [Problem 4], and others. Thus, the RITG enables us to manage all of the scattered information created through research activities in a common environment and organize the information based on its classification (a solution to problem 1). Moreover, visualization through an RITG enables us to easily see the transitions among various types of research information (a solution to problem 2). Management of the research information with individual viewpoints using personal fields enables each researcher to obtain separate information while considering the overall development of the research (a solution to problem 3). Solving these three problems makes it possible to manage the research transitions according to personal viewpoints. Furthermore, classification of the nodes and edges with colors and shapes helps others understand the existing information classifications. It also helps researchers communicate their own valuable information and problems to others.

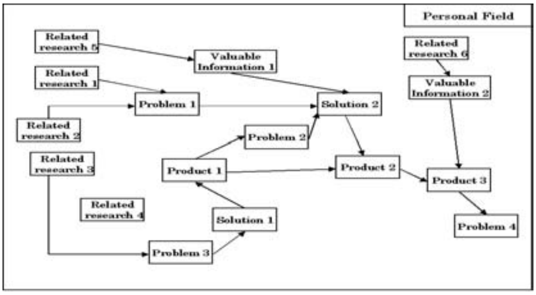


Fig. 1. Example of an RITG

4 Converting an RITG to a Common View

The RITG proposed above has the merit that research transition management can be done through free arrangement of the nodes. However, the feature that allows each researcher's RITG to reflect the researcher's own viewpoints and way of thinking sometimes hampers others' understanding of it since researchers each have their own perspectives. Such a problem hinders information sharing among members of an organization. Therefore, we have developed a method to enhance others' understanding of the research transitions by providing a common view.

To extract a common view, we had the same fifteen subjects as above complete a questionnaire. Before completing the questionnaire, each subject was asked to create an RITG from examples we provided and his own research transitions based on the classifications shown in Tables 1 and 2. The questionnaire results revealed the common views. We defined the following constraints for drawing a common view based on the revealed tendencies.

Constraint 1: the x-axis of the plane in which an RITG is drawn indicates the time sequence (from left to right).

Constraint 2: the field is divided into the number of existing classification areas in the up and down directions, and these divisions are used as the y-axis to arrange nodes.

Constraint 3: the length and breadth of the field is changeable so that nodes can be arranged without overlap.

Fig. 2 shows a common view that satisfies these constraints converted from the RITG based on personal viewpoints. The field is divided into own product, solution, idea, valuable information, problem, detailed information, others, related research and others' product from top to bottom based on Constraint 2. This order is the result of the basic policy that information which many times transits to its own research should be first.

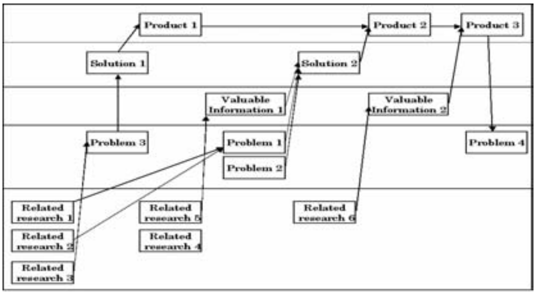


Fig. 2. Common view of the RITG shown in Fig. 1

When an RITG such as that shown in Fig. 1 is used, the rules for drawing the graph are difficult to grasp. If a researcher tries to understand such a graph based on his personal viewpoints, the node positions will often differ from his expectation. It will be extremely difficult for researchers to discover information of interest from the mass of information arranged according to an unknown rule. Conversion into a common view enables others to grasp research transitions based on unified viewpoints. It also enables an understanding of other researchers' current situation and the information available from them. Thus, the common view makes it possible for members of a research organization to exchange information and share valuable information (a solution to problem 4).

The current way to convert an RITG to a common view is still not always easy to understand for all researchers. While it is possible to provide a common view through many different combinations of drawing constraints, in this paper we discuss just one pattern. Increasing the number of patterns will further promote the sharing of research information.

5 Modeling the RITG

The RITG is defined as follows to enable processing by computer.

Definition 1. *Research information transition graph G consists of seven elements:* $G = (V, kindV, timeV, ownV, E, kindE, \pi)$.
V is the finite set of research information. $v(\in V)$ is a vertex of the graph and expresses research information. kindV is an attribute that accompanies the vertex and expresses the classification of the research information. kindV is defined as kindV = {Related research, Product, Valuable information, Problem, Solution,Idea, Detailed information, Other information} The classification of research information $v(\in V)$ is written as $kindV(v)$. timeV is an attribute that accompanies the research information and expresses the date and time of the research information creation. The date and time at which the research information $v(\in V)$ is created is written as $timeV(v)$. ownV is an attribute that accompanies a vertex of the graph and expresses the author of the research information. ownV is defined as ownV = {Name of author}. The author of research information $v(\in V)$ is written as ownV(v). E is the finite set of transition. Transition $e(\in E)$ expresses the transition defined for a research information pair. kindE is an attribute that accompanies an edge of the graph and expresses the transition classification. kindE is defined as kindE = {Update&progress, Origin, Application}. The classification of transition $e(\in E)$ is written as $kindE(e)$. π is the coordinates function of research information $v(\in V)$ defined as $(V \rightarrow R^2)$. The coordinates of research information v are written as $\pi(v) = (\pi_x(v), \pi_y(v))$.

In this research, we let the x and y coordinates start, respectively, from the left and from the top. To construct an RITG, all attributes mentioned above must be given in addition to the research information. When constructing an RITG based on personal viewpoints, coordinates function π can be given freely by each researcher according to his or her viewpoints. When an RITG is converted to common view, the x and y coor-dinates of π are given as the vertex based on the attribute values of the date and time and the attribute of the research information classification, respectively.

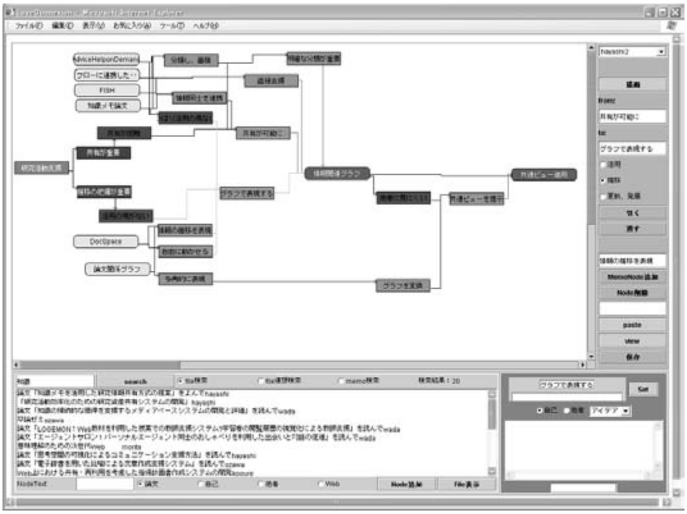


Fig. 3. User interface of the support system using RITGs

6 A Support System for Research Activities

We have developed a system to support research activities which uses RITGs. We used Perl, Java/Swing, and MySQL as development languages and the DBMS. The user interface of the system was implemented on a web browser as a Java applet. This system provides functions for managing, acquiring, and sharing research information. The cooperative work enabled by these functions supports research activities synthetically by enabling both personal management of research transitions and information sharing among members of the organization. Fig. 3 shows the system's user interface.

The function for managing research information enables users to freely arrange the various types of information created through research activities as the nodes of a graph with their classifications and visually express its transitions by using edges to show relationships between the nodes. Even if a user moves any nodes connected with transitions, edges follow the nodes automatically. Moreover, users can refer to and edit the contents of a node by double clicks. When users terminate the system, their RITGs including drawing positions are stored into a database.

The function for acquiring research information enables members of the organization to search through their colleagues' information. Currently, searching using keys such as author name, title, keyword, date of creation and content similarity can be done. Retrieval targets all of the information within the organization. Users can refer to the contents of information obtained by searching. It is also possible to refer to the RITG of the author of obtained information. Users can also arrange obtained information in their own RITG. In such a case, the link information is stored. Conversely, researchers can provide comments and advice to others based on reference to the obtained information. Such comments are displayed in the others field and the receivers can adopt them in their own RITG.

The function for sharing research information enables users to unify their viewpoint for arranging nodes by converting others' RITG into a common view in which nodes and edges follow an orderly arrangement based on their classification and the date and time of creation. Users can also provide information to others' RITGs. These capabilities will promote mutual exchange of research information among members of an organization. Furthermore, users can understand the relevance not only of information obtained by searching but also indirectly related information, its purposes, and others' aims when they refer to one's own RITG by using the functions for acquisition and sharing together.

Thus, the developed functions will promote research activities by supporting users' discovery of valuable new information and ideas while also making the RITGs of others easier to understand.

7 Concluding Remarks

We have described a solution to the trade-off between the need to manage research information with personal viewpoints and the benefits of mutual sharing among colleagues within an organization. Specifically, we have proposed an RITG that enables the management of research information based on personal viewpoints. In addition, we have developed a method for converting an RITG into a common view to enable mutual sharing with unified viewpoints. Finally, a system to support research activities synthetically was developed based on our methods.

In our future work, we plan to develop several common views converted from an RITG because we developed only the one pattern of common view. We also plan to evaluate the system in actual use.

References

1. Umeda K., Yasuda, T., Yokoi, S.: Proposal of a Research Information Sharing System Utilizing Knowledge-Memos. IPSJ J., 42-11 (2001) 2562–2571
2. Smolnik, S., Erdmann, I.: Visual Navigation of Distributed Knowledge Structures in Groupware-Based Organizational Memories. Proc. IEEE Sixth International Conference on Information Visualization (IV'02), (2002) 353–360
3. Ahlers, J., Weimer, H.: Challenges in Interactive Visualization for Knowledge Management. Proc. IEEE Sixth International Conference on Information Visualization (IV'02), (2002) 367–374
4. Tanabe, S., Oyobe, K., Sunaoka, N., Yokoyama, S., Miyadera, Y.: A Visualization System of Relationships among Papers Based on the Graph Drawing Problem. Proc. IEEE Sixth International Conference on Information Visualization (IV'02), (2002) 202–210
5. Guillermo Rodriguez-Ortiz: Knowledge Management and Quality Certification in a Research and Development Environment. Proc. Fourth Mexican International Conference on Computer Science, IEEECS Press, (2003) 89–94
6. Nakakoji, K., Yamamoto, Y.: Knowledgen Interaction Design for Creative Knowledge Work. Trans. JSAI, 19-2, (2004) 154-165

The Learning System of Shinshu University Graduate School of Science and Technology on the Internet

Hisayoshi Kunimune[1], Masaaki Niimura[1], Katsumi Wasaki[2], Yasushi Fuwa[2], Yasunari Shidama[1], and Yatsuka Nakamura[1]

[1] Department of Information Engineering, Faculty of Engineering, Shinshu University, 4-17-1, Wakasato, Nagano City, Nagano Pref., 380-8553 Japan
{kunimune,niimura,shidama,ynakamur}@cs.shinshu-u.ac.jp
[2] Graduate School of Science and Technology, Shinshu University, 4-17-1, Wakasato, Nagano City, Nagano Pref., 380-8553 Japan
{wasaki,fuwa}@cs.shinshu-u.ac.jp

Abstract. Shinshu University Graduate School of Science and Technology on the Internet (SUGSI) is established in 2002 as the first Internet University in Japan. In SUGSI, students can learn every lecture using a CAI system featuring drills on the web, and get supervising about a master's thesis from faculty adviser via network. Therefore, students can complete a master course and get a master's degree without commuting to the school. We manage SUGSI on day school system, and we developed some student support systems as well as learning contents with CAI. In this paper, we mention about the learning system of SUGSI: its CAI contents, a student management system, and the profile of enrolled students such as their age structure and their learning style.

1 Introduction

In April 2002, Shinshu University, Graduate School of Science and Technology on the Internet (SUGSI) was formed as the first Internet university program in Japan and 225 students have enrolled in the first three years of its offering. Students of SUGSI can take lectures on the Internet and earn credits toward a master's degree.

Creating web-based lectures is easy. However, conferring a master's degree with only web-based teaching is not because many problems are encountered in trying to satisfy the Japanese government's requirements. We develop and improve many systems to operate SUGSI to satisfy requirements from the government, lecturers, and students. In this paper, we introduce the systems and measures used in operating the first Internet university in Japan.

2 Learning Processes in SUGSI

2.1 Overview of the Learning Processes

The students of SUGSI can learn with course materials and examinations on the Internet, and can exchange ideas with teachers and other students via email and

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1296–1302, 2005.

BBS. Additionally, the students must study and submit their master's thesis to get the master's degree.

There is no difference between on-site master students and the students of SUGSI, except they learn and study on the Internet or on-site. In fact, on-site students and the students of SUGSI passed the same entrance examination and are treated in the same way on paper.

2.2 Contents for e-Learning

In March 2001, the Japanese government eased standards for the establishment of universities. According to this revision, all graduate schools in Japan can offer their courses with e-Learning lectures that provide the following:

- An integration of various information using texts, sound, pictures, movies, etc.
- Examinations that are corrected by a lecturer and Q&A in each lesson
- Some ways for students to exchange ideas with others

We provide the following contents for each e-Learning lecture to meet the requirement for students.

- Course texts with multimedia
- Web-based examinations
- BBS for the exchange of ideas with others

Students can learn course materials and acquire credits by using these systems. Anyone can browse the above learning contents and the information of SUGSI from the portal site (`http://sugsi.jp/`).

2.3 Course Texts with Multimedia

Currently, SUGSI offers 40 lectures on the Internet. Lecturers make texts of their courses using multimedia such as graphics, animations, and videos (Fig. 1 and Fig. 2). That means there are no violations of copyright, so we impose no limitations on browsing texts in SUGSI. From time to time, students and sometimes people who are not our students point out mistakes in our texts. In such cases, we can easily update the material and can keep the quality of texts very high. It is very important to provide high-quality course texts for students.

2.4 Web-Based Examinations

We provide several types of web-based examinations in every section of a course text based on the contents of the section. For instance, we offer drill type examinations for knowledge acquisition (Fig. 3), exercises of programming language (Fig. 4), and a form for the submissions of electronic reports (Fig. 5).

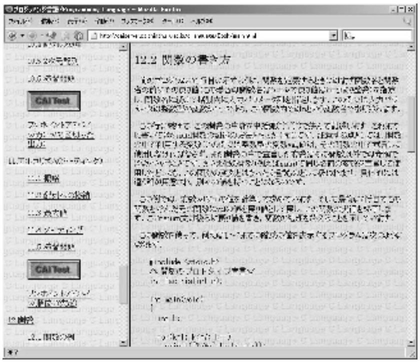


Fig. 1. Course texts

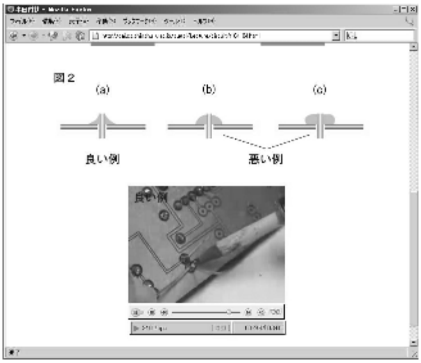


Fig. 2. Course texts with multimedia

In drill type examinations, the drill system selects questions from a pool of questions in a random order and students can repeatedly take the examination. Hereby, students acquire knowledge from these drills[1, 2].

In the exercises of programming language, students write a program in advance and fill in their program source code on a prescribed form. The web server compiles and runs the program posted on the form. Students can check the results of their programs on a web browser.

Fig. 3. Drill type examinations

Fig. 4. Exercise of programming languages

Fig. 5. Form for submission of electronic reports

For report writing assignments, we prepare a system for managing the submission of electronic reports. Students can hand in their electronic reports on a web browser with this system.

All of the above-mentioned examination systems have a database to record students' progress. When a student passes an examination, the examination system commits some information such as "ID number", "time elapsed", "number of tries before passing", etc. to the database. A lecturer can know the learning progression of students only from these data. Therefore, these data of progression are very important for both the lecturer and students.

When students take an examination, a subsequent section of the course text sometimes includes answers or hints to questions of the examination. Such situations impede the learning of students. To solve such situations, we develop a system to present texts based on the learning progression of each student. This system generates passwords from seeds consisting of the student ID, course ID, and a region number assigned to specific sections of the text by the lecturer and students can read texts only for passed sections. When students pass an examination, the examination system generates the password to present or unlock texts of the next section. Students must log in to this system with IDs and passwords generated by this system and they can read texts of the next section.

For instance, when students have not taken any examinations in the course and log in with no passwords, they can read only the first section of the text as shown in Fig. 6. After they pass the first examination and log in with the newly supplied passwords, they will be able to read the next section such as in Fig. 7.

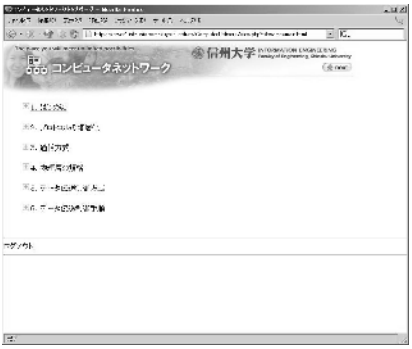

信州大学 INFORMATION ENGINEERING
Faculty of Engineering, Shinshu University
コンピュータネットワーク
ログアウト

Fig. 6. Text before taking examinations

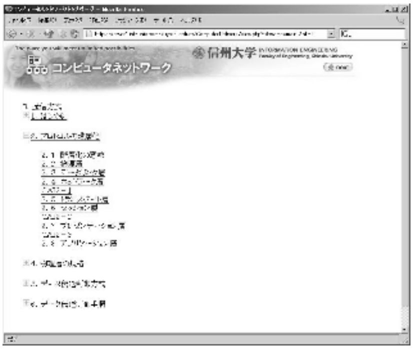

信州大学 INFORMATION ENGINEERING
Faculty of Engineering, Shinshu University
コンピュータネットワーク

Fig. 7. Text after passing the first examination

2.5 BBS

We prepare a number of electronic bulletin boards for exchanging ideas and posting notices about school affairs as follows:

- Information from the university (only lecturers can post)
- Questions and answers concerning SUGSI (anyone can post)
- Questions and answers concerning lectures (anyone can post)

In addition to these boards, each lecture has a BBS to exchange ideas between lecturers and students. Students and lecturers actively use these BBS. In this way, we believe we can respond to the needs of students with these systems.

3 Analysis of Students

Since SUGSI was opened in 2002, 81, 73, and 71 students enrolled in 2002, 2003, and 2004 respectively. We show some analysis results of the students. Fig. 8 shows the age composition and Fig. 9 shows the job composition of students. These figures show that almost 90 percent of the students are between their 20's to 40's and over 80 percent of the students are workers.

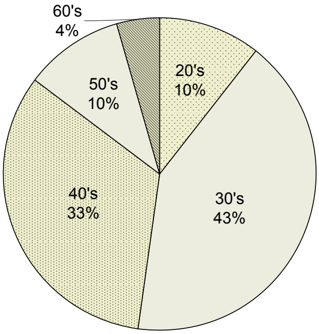


Fig. 8. Age composition of students

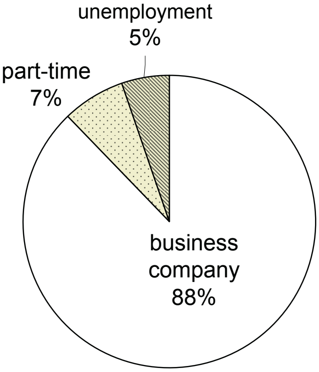


Fig. 9. Job composition of students

Fig. 10 shows the number of students who have passed examinations sorted by hour in weekdays and holidays. This figure shows that a relatively large number of students take examinations and there are students taking examinations at all hours. Therefore, it is very important to make sure our systems work at all times.

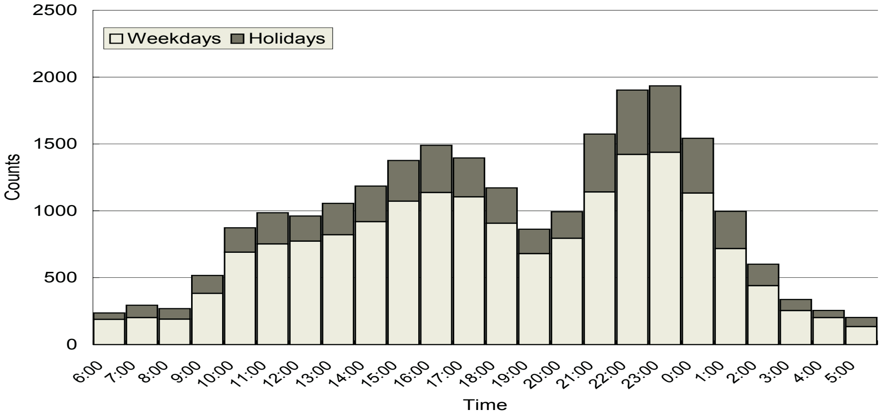


Fig. 10. Hourly distribution of students passing examinations

4 Servers and Networks

We mentioned that lecturers can know the learning progression of students only from the data committed to databases by the examination systems in Section 2. Additionally, we mentioned that it is very important to make sure the system of servers and networks are working at all times in Section 3. We consider and implement measures to protect progression data and to increase the reliability of the servers and networks as described in this section.

Fig. 11 shows the composition of servers and networks of SUGSI. Servers in this figure provide services as follows.

- The web server hosts learning texts and BBS services.
- The VoD server provides streaming videos.
- Examination servers generate examinations and house the learning progression data of the examinations provided on each server.
- The progression database server holds the learning progression data of all examinations.

We place web and progression database servers in both the networks of Shinshu University and two outside ISPs. Such a redundant configuration of servers ensures these servers work at all times and fulfills the demand of students. The mirrored web server at the ISP1 makes a mirror of data of the web server in Shinshu University once a day.

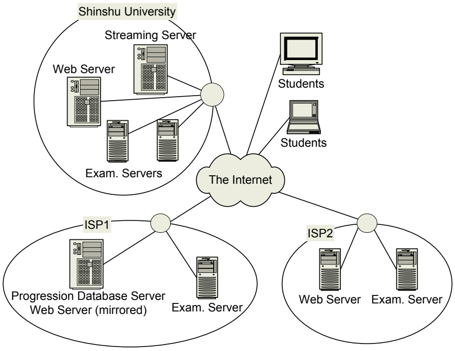


Fig. 11. Composition of SUGSI servers and networks

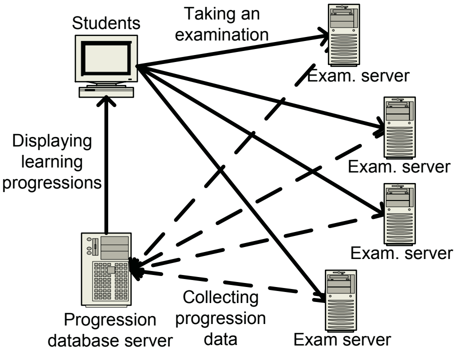

Students
Taking an examination
Exam. server
Exam. server
Exam. server
Exam server
Displaying learning progressions
Progression database server
Collecting progression data

Fig. 12. Flow of progression data

Fig. 12 shows the flow of progression data. Lecturers develop their own examination systems to fit the characteristics of each lecture on a server, and examination systems hold progression data on each server. The progression database server collects progression data from examination servers once a day. Students as well as lecturers can confirm the progress of learning at any time. If progression data on the database server or examination servers are lost, we can restore the data from the examination servers or the database server.

Fig. 13 shows a display of progression data for a student. In this page, students can confirm examinations passed from all lectures. This page helps students to plan their curriculum studies. Fig. 14 shows a page displaying progression data for lecturers. Lecturers can confirm how many examinations each student has passed from this page.

Fig. 13. Page displaying progression data for a student

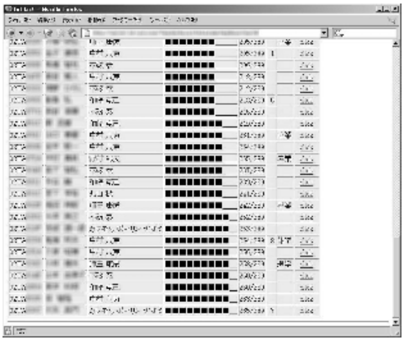


Fig. 14. Page displaying progression data for lecturers

5 Conclusion

From 2002, we have been operating the learning system of SUGSI including its online course texts, BBS, and databases, and have developed many systems to support their operation. In this paper, we mentioned many systems, however, they are just a part of the entire system. We receive many kinds of feedback and requirements via BBS, email, etc. from lecturers and students and based on this input, we improve the systems and develop new systems.

We believe we can satisfy social needs. In fact, 225 students enrolled at SUGSI in the past three years, and over 40 students will enroll in April 2005. Additionally, almost students give very high evaluations for the learning systems of SUGSI. However, several universities will start and many others will prepare to open on the Internet in Japan. Therefore, we must strive to improve SUGSI contents and systems to continuously satisfy the needs of students.

References

1. Fuwa, Y., Shidama, Y., Wasaki, K., and Nakamura, Y.: The Plan of the Shinshu University Internet Graduate School. JSiSE. 19, 2 (2002) 112-117
2. Fuwa, Y., Nakamura, Y., Yamazaki, H., Oshita, S.: Improving University Education using a CAI System on the World Wide Web and its Evaluation. JSiSE. 20, 1 (2003) 27-38

Automatic Generation of Answers Using Solution Network for Mathematical Exercises

Tomoko Kojiri[1], Sachiyo Hosono[2], and Toyohide Watanabe[2]

[1] Information Technology Center, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
kojiri@watanabe.ss.is.nagoya-u.ac.jp
[2] Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
{hosono,watanabe}@watanabe.ss.is.nagoya-u.ac.jp

Abstract. The objective in this paper is to construct a mechanism for generating the answers for mathematical exercises automatically. A mathematical exercise is decomposed into several small exercises that are called sub-exercises. An answer of an exercise is a collection of answers of its sub-exercises. If the inclusive relations among exercises are managed, answers for exercises can be generated easily. In our approach, the inclusive relations of exercises are represented as a solution network. The solution network is structured as a tree. In the solution network, nodes correspond to the exercises and links indicate inclusive relations of exercises. By traversing the solution network from the node corresponds to the target exercise and detecting its sub-exercises, the answer of the target exercise is able to be generated automatically.

1 Introduction

Currently, with the development of network, the learning through the internet is one of the hottest topics [1]. In such an e-learning environment, amounts and types of authoring texts are important. If the authoring texts are not provided sufficiently, learners cannot acquire enough knowledge.

We have been constructed the collaborative learning support environment called HARMONY [2][3]. HARMONY focuses on the collaborative learning of high school mathematical exercises and provides the assistant function which leads the learning group to its learning goal. In high school mathematics, learners understand formulas by tackling many exercises repeatedly. So, many exercises of the same type are required. Currently, each exercise is defined by the authors of HARMONY beforehand, so it needs a lot of work for them to prepare huge amount of exercises. Therefore, the objective in this paper is to construct a mechanism for generating the answers in mathematical exercises automatically.

Hirashima, et al. provided MIPS (Model of Indexing Problem Solving) which describes knowledge of exercise, conditions to apply the solution method, and procedure to be carried out when the solution method is applied [4]. Also, Matsuoka, et al. suggested WEDOM (WBT Drill Exercise Organization Model) in which not only the elements in the exercise but also groups of variables used in the exercise are modeled [5]. These researches focused on modeling the knowledge of rather small area, in which relations and procedures among knowledges are fixed.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1303–1309, 2005.

A mathematical exercise is decomposed into several small exercises that are called *sub-exercises*. The relation between the exercise and its sub-exercise is defined as *inclusive relation*. The sub-exercise also has its sub-exercises and the primitive exercise is defined as a exercise which consists of only one answering step. An answer for an exercise is a collection of answers for all its sub-exercises. In order to generate the answer of the exercise, to specify its sub-exercises is needed.

In our approach, the inclusive relations between exercises are represented as a solution network. The solution network is structured as a tree, in which nodes correspond to the exercises and links indicate inclusive relations of exercises. Namely, an exercise of a child node is a sub-exercise of its parent node. Moreover, the way to describe the answer is defined according to the exercise. By traversing the solution network from the node whose goal is same as the target exercise, the sub-exercises of the target exercise are determined and description for the exercise is generated.

2 Approach

2.1 Characteristics of Mathematical Exercise

The characteristics of the answer for mathematical exercises are as follows:

1. *An exercise consists of several sub-exercises.*
   For example, in Figure 1, the exercise for calculating the minimum value in two dimensional function contains the exercise for deriving the peak. Similarly, the exercise for deriving the peak consists of the exercise for completing square.
2. *Each exercise has its own answering methods.*
   An answer of an exercise consists of several answering steps which correspond to answering methods respectively. In addition, each exercise has its own answering step. For example, in Figure 1, the exercise for calculating the minimum value consists of 4 answering steps and the last answering step "to determine minimum value" is unique to that exercise.
3. *Exercise has several answering paths.*
   Answering path is a path to derive the answer. If there were two answering paths for the same exercise, their sub-exercises are different. For example, in Figure 2, exercise for determining two dimensional function has two answering paths; one consists of sub-exercise for substituting points and the other includes sub-exercises for substituting points and answering simultaneous equation.

2.2 Generating Answer in Mathematical Exercise

To generate an answer for a mathematical exercise consists of two steps. In the first step, answering methods are determined according to the condition given in the exercise. Answering methods have preconditions to be applied. After applying them, particular equations are generated. So, appropriate answering methods whose preconditions satisfy the condition of the exercise and which generate appropriate equations for the next answering step should be selected.

In the second step, the descriptions for the answering methods are generated. Descriptions for the same answering methods are the same. However, variables, such as coefficients and constants, are different. They are determined by the equations given

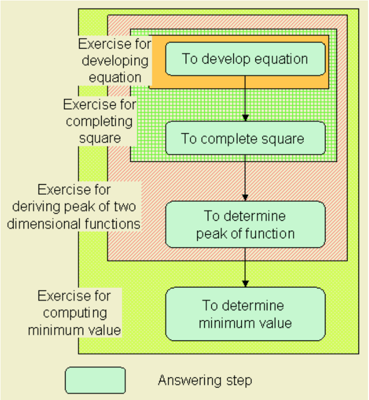

Exercise for developing equation
To develop equation
Exercise for completing square
To complete square
Exercise for deriving peak of two dimensional functions
To determine peak of function
Exercise for computing minimum value
To determine minimum value
Answering step

Fig. 1. Example of exercise I

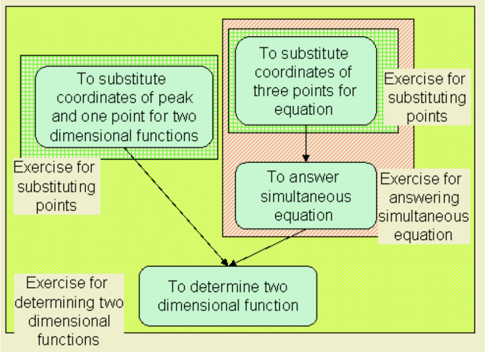

To substitute coordinates of peak and one point for two dimensional functions
Exercise for substituting points
To substitute coordinates of three points for equation
Exercise for substituting points
To answer simultaneous equation
Exercise for answering simultaneous equation
Exercise for determining two dimensional functions
To determine two dimensional function

Fig. 2. Example of exercise II

in the exercise. Therefore, the template for describing each answering method should be prepared and transformed so as to fit to the target exercise.

2.3 Framework

Figure 3 shows the framework of our system. The system consists of two mechanisms: one is to detect answering methods for the exercise and the other is to generate the descriptions for the exercise. The answering methods belong to the exercises and the exercises have inclusive relations. In order to represent the relations of exercises and specify the sub-exercises of the target exercise, the solution network which represents inclusive relation among exercises is introduced. The solution network is structured as a tree whose nodes correspond to the exercises and links indicate inclusive relations. Sub-exercises are arranged in the sub-tree of the node of the target exercise. In addition, answering methods are attached to the exercise. Therefore, by traversing the solution network from the node for the target exercise, the sub-exercises are detected, and hence the answering methods are determined.

The method for describing answering method is the same even if the constants or the coefficients of exercise were different. Therefore, the template of describing answering methods is also attached to the solution network. The description for the target exercise is generated according to the templates of answering methods and the conditions given in the exercise.

3 Solution Network

3.1 Structure

Solution network represents inclusive relations among exercise. Figure 4 shows the concept of our solution network. Nodes represent exercises and links represents the inclusive relations between exercises. Nodes in the sub-tree correspond to its sub-exercises. For example, in Figure 4, answer of the exercise A contains the answer of the exercise B, and answer of exercise B also has the answer of exercise D.

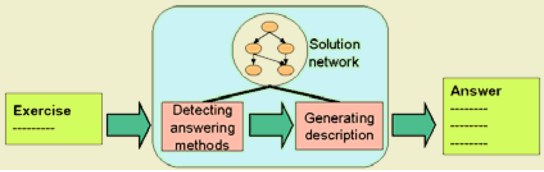


Fig. 3. Framework of system

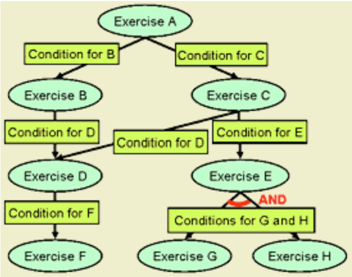


Fig. 4. Concept of solution network

Answering methods specific to exercises are attached to the nodes. Here, we assume the exercise in Figure 5a. The solution network corresponds to the exercises in Figure 5a is shown in Figure 5b. Since exercise B is the sub-exercise of exercise A and exercise C is the sub-exercise of exercise B, the node of exercise C is a child node of exercise B and the node for exercise B is a child node of exercise A. Moreover, the answering methods 1, 2, and 3 belong to nodes for exercise A, B, and C, respectively.

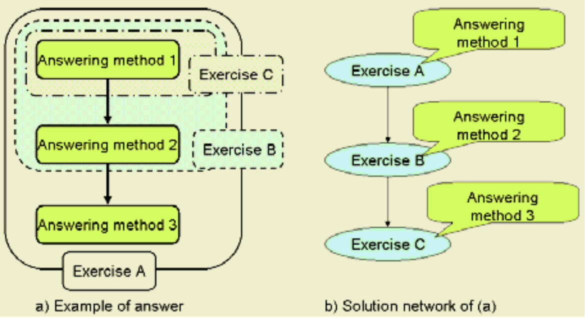


Fig. 5. Example of constructing solution network

On the other hand, there are two kinds of links; AND link and OR link. AND link means that exercise of the parent node includes the sub-exercises in one answering path. On the other hand, OR link indicates that the exercise has several answering paths and one sub-exercise is included in each these answering path.

If the exercise has several answering paths, the appropriate answering path for the target exercise is selected according to its conditions, such as goal and equations. In order to determine appropriate path in the solution network, the conditions for traversing the solution network are attached to links.

3.2 Conditions on Links

Conditions on links are defined for selecting sub-exercises for the exercise. The conditions are described in the predicate form. Table 1 shows some of the predicates that are prepared for the field of two dimensional functions.

Table 1. Examples conditions

• Eq(a, b): a = b • Gt(a,b): a>b • Lt(a,b): a<b • X-Axis(a):x=a • Y-Axis(a):y=a	• Const(a): a is a constant. • Var(a): a is a variable. • Function(f.x): f is a function of variable x. • Equation(e): e is a equation.	• Peak(p): p is a peak • Max(a,b): Maximum value of a is b. • Min(a,b): Minimum value of a is b.

When the target exercise is given, sub-exercises are selected by traversing nodes in the solution network whose links satisfy the condition of the exercise. Following is the algorithm of detecting sub-exercises.

Step1: Find the node which corresponds to the goal of the exercise and set the conditions given in the exercise as a current situation.

Step2: Compare the conditions of links with current situation.

If conditions of link satisfy the current situation, set the child node as a current node and continue step2.

If the conditions of all its links do not satisfy the current situation, go to step3

Step3: Order the traversed nodes from the lower to the upper, output them, and finish traversing procedure.

In the solution network, exercises in the lower level are sub-exercises of the higher level. If the answer of exercise requires more primitive answering methods, such as computation, solution network is traversed into deeper level and many sub-exercises are detected. On the other hand, if the exercise gives equations calculated as its conditions, small number of sub-exercises is specified.

3.3 Descriptions of Answer

The description of the same answering method is the same. Only coefficients and constants are changed according to equations or numbers given in the target exercise. Moreover, the coefficients and the constants are able to be calculated based on the conditions given in the exercise. So, the templates of descriptions and its methods for calculating coefficients and constants are attached to the nodes of solution network.

Table 2. Example of answer template

Condition	Var(x), Const(a), Const(b), Const(e), Const(d), Const(e), Function2(f, x), Equals(f, $a(x-b)^2$+c), Calc(a>0), Ge(x,d), Le(x,e), Calc(e<b)
Description	Since a > 0 for the given two dimensional function, its axis x=b. Since the range of x is d≦x≦e, two dimensional function f is minimum when x = e. $f=a*(e-b)^2+c =a*(g)^2+c = h+c=i$ Therefore, the minimum is I when x = e
Variables	Const(g), Const(h), Const(i), g =e-b, $h=a*(g)^{2,}$, i=h+c
Output	Min(f,i)

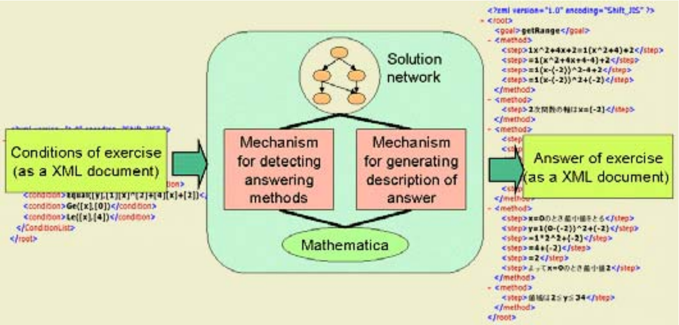
Templates consist of 4 parts. Table 2 shows the example of template. Since the descriptions and the method for calculating the coefficients and constants are different according to the learning situation, conditions of applying that template are defined as

“Condition”. The predicates of expressing the conditions are the same as the predicates on the links. “Description” is the template for describing the answer. The method for calculating the coefficients and constants in the description is prepared in “Variables”. Using the conditions given by the exercise and the calculation methods in “Variables”, the coefficients and constants in the description are determined. The predicates in the field “Output” show the result of applying the answering method. After the answering method has been applied, the indicated predicates are generated.

4 Experiment

4.1 Prototype System

Figure 6 shows the overall framework of our prototype system. Conditions of the exercise written as XML document are input to the system and an answer of the exercise is output as XML documents. Currently, the system generated only one answer which satisfies the condition of the exercise. In the system, the answering methods used in the answer are detected based on the solution network, and then the descriptions for answering methods are generated. The conditions for links and conditions given in the exercise are managed as Java objects. If many conditions are matched to the current learning situation, the one whose number of the conditions is biggest is selected. Mathematica [6] is also used in the system in order to calculate the variables or to compare the numbers.



```
    <condition>Equal([y],[1][x]^[2]+[4][x]+[2])<
    <condition>Ge([x],[0])</condition>
    <condition>Le([x],[4])</condition>
  </ConditionList>
</root>
```

```
<?xml version="1.0" encoding="Shift_JIS" ?>
- <root>
   <goal>getRange</goal>
 - <method>
    <step>1x^2+4x+2=1(x^2+4)+2</step>
    <step>=1(x^2+4x+4-4)+2</step>
    <step>=1(x-(-2))^2-4+2</step>
    <step>=1(x-(-2))^2+(-2)</step>
   </method>
 - <method>
    <step>2次関数の軸はx=[-2]</step>
   </method>
 - <method>
 - <method>
    <step>x=0のとき最小値をとる</step>
    <step>y=1(0-(-2))^2+(-2)</step>
    <step>=1*2^2+(-2)</step>
    <step>=4+(-2)</step>
    <step>=2</step>
    <step>よってx=0のとき最小値2</step>
   </method>
 - <method>
    <step>値域は2≦y≦34</step>
   </method>
</root>
```

Fig. 6. Overall framework of prototype system

4.2 Experimental Result

We have constructed the solution network in two dimensional function and experiment if the system could derive the appropriate answer. The solution network was designed by analyzing answers of 40 exercises in the textbook. Our solution network consists of 19 nodes and 21 links.

25 exercises from different textbook are used in our experiment. The nodes of the answering methods used in those exercises are prepared in the solution network. Table 3 shows the result of the experiment.

The answers of all exercises were generated and most of them were appropriate, because they were described in the textbook as appropriate answers. However, there is one answer which was not appropriate. The reason for not having been generated appropriate answer is that the node which holds appropriate answering method is not prepared in the solution network. If the solution network holds the corresponding node, that appropriate answer may be derived.

Table 3. Result of experiment

Output answer was correct	Output answers were the same as those in the textbook.	24	25
	Output answer was correct, but not the same as that in the textbook.	1	

5 Conclusion

In this paper, the automatic generation of the answer for the mathematical exercise based on the solution network was proposed. Based on the experimental result, the system could generate the appropriate answer from the conditions of exercises, if the nodes of the answering methods are prepared in the solution network.

For our future work, the solution network should be extended for the purpose of generating answers for more complex exercises which contains answering methods in various fields. In order to cope with more complex exercises, not only the solution networks for the different field, but also the conceptual network which connects concepts that are common to the different fields are needed.

References

1. H. H. Adelsberger, B. Collis, and J. M. Pawlowski (eds.): Handbook on Information Technologies for Education and Training, Springer- Verlag (2002).
2. T. Kojiri and T. Watanabe: Fundamental Mechanism for Activating Discussion in Collaborative Learning or Mathematical Exercise – Detecting Answering Viewpoint Based on Diagrams –, Proc. of CATE 2003, pp.426-431 (2003).
3. T. Kojiri and T. Watanabe: Diagrammatic Advices in Mathematical Exercises, Proc. of ICCE 2003, pp. 293-296 (2003).
4. T. Hirashima, T. Umeda, and A. Takeuchi: A Intelligent Problem-Authoring Environment on the Web, Proc. of ICCE 2001 pp.459-466 (2001).
5. H. Matsuoka, K. Kanenishi, H. Mitsuhara, K. Matsuura, and Y. Yano: The Creation Support of Knowledge Base for Exercise Generation System, Technical Report of IEICE, Vol. 103, No. 697, pp. 179-184 (2003) [in Japanese].
6. Inc. Wolfram Research: Mathematica, http://www.wolfram.com/

Assisting Construction of Meta-cognitive Competence by Scaffolding Discovery of Plans in Problem-Solving

Kohji Itoh, Eisuke Mihara, Kenji Hasegawa, Masahiro Fujii, and Makoto Itami

Tokyo University of Science
2641 Noda, 278-8510Japan
itoh@te.noda.tus.ac.jp

Abstract. In our collaborative environment for learning by problem-solving, whose implementation so far will be briefly described, we provide the learners with an environment in which they edit plans for solving given problems by selecting and connecting problem types provided by the system. In the problem types, on which we first discuss, context-dependent plans consisting of problem types are embedded. The learners should recursively edit plans. Embedded plans, however, are provided, when they get deadlock. A library of sample problems are also provided so as to be examined by the learners to learn what are and how are used the plans and the problem types in concern. Finally we propose, as learning control to realize construction of meta-cognitive competence, to make use of the library of sample problems with solution to scaffold discovering plans for the problem given to the learners, through retrieving and providing pertinent sample problems using, as indices, the problem types and plans as well as annotations for bridging problem types or plans.

1 Introduction

We have been implementing a framework of explorative environment of collaborative learning by problem-solving[1][2], providing the learners with an environment in which they edit plans for solving given problems by selecting and connecting problem types provided by the system. In the problem types, context-dependent plans consisting of problem types are embedded in place of declarative descriptions of knowledge repertoire. The learners should recursively edit these plans which are provided, however, when they get deadlock.

The reason why we employ this framework first comes from the pragmatic point of view saying that merely memorizing a collection of basic knowledge would not work for solving problems (e.g. Polya [3], Clancey [4]) Planning strategies on what knowledge could be applied when and how is of the first importance.

Secondly, the usage of strategy and knowledge largely depends on the context of the problem in which it is used, and this dependency has come to be known unable to be dealt with by simple combination of independently generalized context, strategy and knowledge[5].

One of the rare examples of ITS implemeneted from planning point of view was GUIDON [4] made by Clancey which assisted medical students of diagnosis of blood infectious diseases.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1310–1316, 2005.

In this paper we first discuss backgrounds on which we provide the learners with problem types in which embedded are context-dependent plans. We briefly describe the implementation of the collaborative learning environment so far made.

Finally we propose, as learning control to realize construction of meta-cognitive competence, to make use of the library of sample problems with solution to scaffold discovering plans for the problem given to the learners, through retrieving and providing pertinent sample problems using, as indices, the problem types and plans as well as annotations for bridging problem types or plans.

2 Problem Types and Plans

Consider, for instance, the useful strategy "divide and conquer". In obtaining the volume of a given solid body the strategy may be applied by separating the body into a number of parts, obtaining the volumes of the components and adding them. In solving statics or dynamics problems in mechanics, force vectors could be divided, according to the strategy, into components appropriate for solving balance or obtaining motion equations. For obtaining the current of a branch in an electrical circuit, it is often the best way to separate the circuit into 2 parts cutting the branch, convert each part into an equivalent simple circuit, reconnect them and solve the simplified circuit to obtain the current in concern. In drawing such geometric figures as satisfying given conditions, it is suggested, as in [3], to decompose or transform the conditions into a few different conditions each of which should be satisfied by points forming a trajectory and based on their cross points the desired figure could be obtained.

It is rather obvious that a mere collection of basic knowledge units would not work for solving problems and planning strategies are of utmost importance. The point of the importance, however, lies in that each of these applications of the same strategy requires different interpretations depending upon the context of the problem types and different knowledge is necessary to be used in different types of problems.

Conversely, different problem types with appropriate plan options should be provided for dealing with different worlds.

G.Polya suggests in [3] such heuristics for solving mathematical problems as most general as they could be. He, however, stresses importance of experience of discovery of plans in actually solving problems. Therefore the system for developing ability of planning should provide an environment that could scaffold planning from the realistic points of view. i.e. context dependency of planning.

These are why we introduced problem types with plan repertoires.

Now the process of solving a problem can be viewed as selecting such an answer we call "ASKED" as satisfying the constraint we call "GIVEN", from the repertoire of the candidates specified in relation to "GIVEN".

GIVEN specifies the problem type as well as the category and the constraints/ conditions directly/indirectly imposed on ASKED.

Observing that any problem-solving results in a knowledge unit either of the types "(a) If A holds then B holds" or "(b) Under context C, A is equivalent to B" , one can obtain the following 8 categories of abstract problem types.

(a) A is GIVEN and B is ASKED: e.g. analyzing attribute values, causality,
B is GIVEN and A is ASKED: e.g. design, diagnostic, drawing figures.
A,B are GIVEN and ASKED why: proof of an implication.
(b) C,A are given and B are asked: e.g. solutions of equations.
C,A,B are given and ASKED why: e.g. proving solutions satisfy equations.
Facts are GIVEN and ASKED is a knowledge unit supported by GIVEN
: e.g. inductive or abductive knowledge construction.
Facts are GIVEN and ASKED is a concept supported by GIVEN
e.g. introduction of concepts.
GIVEN a problem and ASKED is planning for solving

And a problem types is defined by combination of an abstract problem type and a problem context so as the plan options being defined as connections of problem types are properly introduced according to their contextual nature.

Because the objective of introducing problem types is to facilitate the problem solving agents, e.g. learners, to discover a pertinent problem solving procedure by examination of the problem description, a problem type is in effect a cluster of plans with a common target context, and its explication or narration should serve for the agents rapidly to be trained to identify and differentiate their usage in planning at each recursive stage of designing plans for the sub-problem of the stage.

Therefore, the principle of defining problem types is balance between reusability and differentiability.

A problem type as a component in a plan may again requires plans for solving or may have explicit usage of a declarative or procedural knowledge unit which we call a leaf problem type.

We show some examples of problem types in Appendix. As for "Annotations" and "sample problems" , see section 4.

As you see in the examples, we observe, in human problem solving processes, ubiquitous use of plans consisting of sub-problem types which may have also plans alike, while any systematic study of repertoires of specific problem types and structures of plans, however, has almost never been made in Artificial Intelligence field. It may have been that problem-solving was considered nothing more than searching for connection of the basic knowledge units.

3 A Collaborative Environment for Learning by Problem-Solving

In our implementation of the Collaborative Environment for Learning by Problem-Solving we provide the learners with a menu of problem types in which embedded are context-dependent plans.

A library of sample problems are also provided and at least a couple of sample problems should be retrieved from any of the plans of the problem types which is used in solving them, and they can be examined by the learners to learn what are and how are used the plans and the problem types in concern.

Each of the problems types is provided with assistance for planning and executing solution. In the final version of the system to come, for all of the sample problems, as long as the solution follows the embedded plans, the problem types the solution plan visits are to be equipped with model-tracing type execution assistances.

The learners are given problems the system selected from the library of sample problems according to the learning control yet to be implemented. And they are requested to describe the problem as to what are given and what are asked, which will help them understand the problem, as G. Polya suggests. They are invited to design a plan for solving the problem by selecting problem types from the menu and connecting them on the plan map view, and required to follow the plan, redoing if they think it deserves. When learners get deadlock in planning or executing solution of a subproblem of certain problem type , they can ask the system to unfold the embedded plan or to provide sample problems which use the problem type.

The learners will be able to collaborate on the collaborative workbench environment in all of these activities, in the near future, working together on the collaborative plan map and the execution assistance window with dialogue in voice and marking focus by handwriting and voice recognition, all to be recorded along with the sequence of the activities in synchronism. Collaborative reflection will be made possible by retrieving mark-ups and replaying the record there-around.

4 Scaffolding Discovery of Plans

As G.Polya repeatedly suggests, discovery of plans in problem solving takes place through guessing based on analogy between the problem to be solved and other problems already solved. We observed that we see analogy between the problems through the type of the problem, the plan or some issues observed in the sub-problems comprised in the plan.

In order to enhance the possibility of analogy taking place, we propose to optionally add annotations either to the problem types or to the plans in our library of problem types as you see some examples in the appendix. The annotation designates the problem type or the plan is related to a problem type or plan. The purpose of the annotation is solely to relate problems which remain isolated otherwise.

Now we propose learning control "scaffolding discovery of plans" in order to realize construction of meta-cognitive competence, i.e. the ability of discovering plans for problem-solving. It will be implemented as the core learning control mechanism in our Collaborative Environment for Learning by Problem-Solving.

For example, suppose you are given a electrostatic conductor system and you must solve for the voltages given the charges of the conductors. As you see in the appendix, the problem type E1 dealing with the electrostatic conductor system is annotated that it belongs to problem type E solving for attributes of a system linearly related to given causes. If you request the system to look for other problem type samples also belonging to E. Problem type A is found to have such annotation and a sample problem AI of the type A is provided with solution. You will concretely see correspondence between (charge, voltage) and (source voltage, loop current) pairs and solving linear simultaneous equations as plan. Thus you will be given hints on solving the (charge voltage) causality

Next example is on a mechanics sample problem BI for solving for the motion of a pendulum point mass on a smooth slope. The system will advice to examine and compare problems BII, BIII, BIV belonging to the same problem type B. Application of plan B2 in solving linear motion problem BII will give the fundamental idea of using the principle of energy and work. But you see in BIII the same principle also can be applied to a curved motion. You will have got a hint of using the principle to your problem of pendulum on the slope. Yet question is how about the work which may be done by the tension force of the cord. Comparison between the descriptions of solving problem BIV and BIII answers the question. For in case of motion on a slope with friction, the frictional force does non-zero work because it is to the direction of the movement and in case of jet coaster, the work done by the normal reactive force is always zero because it is always perpendicular to the direction of the movement.

The third example also concerns hints given by sample problems belonging to the same problem type as the given problem does. Thus in planning for problem CI: obtaining the volume of a frustum, use of Plan C3 of subtraction principle in solving problem CII gives hints.

5 Conclusion

In our collaborative environment for learning by problem-solving, whose implementation so far was briefly described, we provide the learners with an environment in which they edit plans for solving given problems by selecting and connecting problem types provided by the system. In the problem types, on which we first discussed, context-dependent plans consisting of problem types are embedded. The learners should recursively edit plans. Embedded plans, however, are provided, when they get deadlock. Finally we proposed, as learning control to realize construction of meta-cognitive competence, to make use of the library of sample problems with solution to scaffold discovering plans for the problem given to the learners, through retrieving and providing pertinent sample problems using, as indices, the problem types and plans as well as annotations for bridging problem types or plans. Developing learning control based on open learner modeling is a task to be pursued.

References

1. Fujihira,M.,Kawamura,T.,Kawakami,K.,Itami,M.,Itoh,K: CAFEKS: An Atchitecture for Evolutional Development of Interactive Learning Environmnt with Coached Problem- Solving, Proc. ICCE99, vol.1 (1999) 915-922
2. Kawkami,K.,Watanabe,T.,Tateno,M.,Tabaru,Y.,Itami,M.,Itoh,K.: Learning by Problem-Solving Assisted of Planning and Retrieval of Sample Problems with Solutions Indexed by Structured Solution Plan, Japan Journal of Educational Technology,vol.25,no.2 (2001)
3. Polya,G.: Mathematical Discovery, Vol.1-2, John Wiley & Sons (1962)
4. Clancey,W.J.: A Knowledge-Based Tutoring: The GUIDON Program, MIT Press (1987)
5. Fensel,D.: Methods of Problem-solving, Springer-Verlag (1998)

Appendix: Examples of Problem Types and Plans

ProbType A: ASKED currents or voltages of a GIVEN DC electric circuit
(Annotation: ProbType E: ASKED is attributes of GIVEN system which
are linearly related with GIVEN causes .
Sample problem type E1 : ASKED are voltages of GIVEN conductors

```
    forming an electrostatic system with GIVEN charges of the conductors.)
  Plan A: Let {E} be null
    ProbTypeA1: Search for an equation comprising a member of the ASKED
        either applying Ohm's Law, Kichhoff's Voltage Law or Kirchhoff's
        Current Law and add it to {E} and add the new unknowns to ASKED.
        If {E} can be solved for ASKED, then it ends, else ProbType A1 .
  Plan A2:
    ProbType : Transform GIVEN DC circuit into an equivalent circuit using
        serial  resistor /parallel resistor /Tevnin/Norton transformations.
    ProbType A  .
    Recapture I if necessary.
  Plan A3:
    ProbType :  Find closed loops ad hoc being independent and sufficient.
    ProbType :  Obtain the loop equations using Kichhoff's voltage equations.
    ProbType:  Solve the simultaneous linear equations for LOOP CURRENTs.
    ProbType:  Obtain ASKED in terms of the LOOP CURRENTs.

ProbType B:  ASKED is motion of a GIVEN point mass in the gravity field
  Sample problem BI
    ASKED is the pendulum motion velocity v, as a function of position x , of
    a GIVEN point mass on a GIVEN smooth slope in the gravity field.
  Sample problem BII
    ASKED is the height reachable of a GIVENpoint mass vertically projected
    with a GIVEN velocity.
   Sample problem BIII
    ASKED is the velocity of a GIVEN train (seen as point mass) at any point
    on the rail of the GIVEN frictionless jet coaster.
  Sample problem BIV
    ASKED is the reachable height of a GIVEN point mass on a GIVEN
    slope with friction.
  Plan B1:
    ProbType :Set up the coordinate system.
    PronTypeB11
        ProbType: Assuming constraining forces, decompose the forces along
        the coordinate axes, express constraints in equations.
        ProbType: Obtain the differential equations for x(t) using Newton's law
        ProbType: Solve for motion using the equations.
        If the result meet GIVEN conditions, then ends,
        else recursively ProbType B11
  Plan B2:
    ProbType: Set up the coordinate system.
    PorbType: Obtain the kinetic energy of the mass point in terms of v
    ProbType: Obtain the gravity potential energy of the mass point in terms of x
    ProbType: Obtain work done by the force other than the gravity
    ProbType: Express the principle of energy and work done in a equation.
    PorbType: Solve the equation for v as function of x.
```

ProbType C : ASKED is the volume of a GIVEN solid figure.

- Sample problem CI: ASKED is the volume of a GIVEN rectangular frustum with its height, the lengths of the upper and bottom edge.
- Sample problem CII: ASKED is the volume of a GIVEN rectangular parallelepiped with a corner deleted of a GIVEN figure similar to the whole.
- Plan C1
  - ProbType : Directly calculate the ASKED volume of GIVEN figure.
- Plan C2
  - ProbType: Divide the GIVEN solid figure into regular parts {Fi}.
  - ProbType: Obtain data necessary to calculate the volumes {Vi of Fi } .
  - ProbType: Calculate {Vi}
  - ProbType: Add {Vi}to obtain the ASKED
- Plan C3
  - ProbType: Represent the GIVEN solid figure by difference of figures F1 and F2.
  - ProbType: Obtain data necessary to calculate the volumes {Vi of Fi } .
  - ProbType: Calculate {Vi}
  - ProbType: Subtract V2 from V1 to obtain the ASKED

Knowledge Level Design Support for SCORM2004-Conformed Learning Contents – Ontological Consideration on Platforms for Intelligent Educational Systems

Mitsuru Ikeda and Yusuke Hayashi

School of Knowledge Science, Japan Advances Institute of Science and Technology
{ikeda,yusuke}@jaist.ac.jp

Abstract. This paper discusses learning contents design from the viewpoint of knowledge level and symbol level. The purpose of study is to develop a foundation for share and reuse of IESs on a global platform. SCORM2004 is becoming de facto standard so we choice it as the basis of the platform. On the platform we aim to build an environment for authors to clarify pedagogical meaning of learning contents based an ontology for IESs. This approach will allow us to share and reuse academic and technical expertise in the field of AIED research on common platform.

1 Introduction

In the research area of designing instructional systems, we have been aiming at a paradigm shift from "Story board representations of instructional material to more powerful knowledge based representation"[Murray 98]. Major benefit of the knowledge based representation is the realization of highly adaptive instruction with the integrated knowledge bases of learning domain, teaching strategies and learner models. However, building the knowledge bases still requires a significant cost. In order to bring about a solution for these issues, many efforts have been carried out in our IES community.

The thought of *Knowledge Level* by Newell [Newell 82] is seen to be value of designing intelligent systems. The Knowledge Level is a level of description of the knowledge of intelligent systems and the symbol level is one produces the intelligent behaviour based on the knowledge level description. If an intelligent system is high quality, the system is designed in a harmonious balance between the knowledge level and the symbol level.

Issues discussed in this paper are that what is support for structuring well-organized knowledge for intelligent educational systems and that what is adequacy of mechanisms for emerging intelligent behaviour based on the knowledge.

The authors think the keys to the issues are ontological engineering in terms of the former and scalability and interoperability, which are flowing from standards for e-learning, in terms of later.

This paper shows an advanced stage of our research activities on ontology-aware authoring tool [Hayashi 04] but the results are only in early stage. Based on the study,

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1317–1323, 2005.

this paper discusses analysis of SCORM2004[ADL 04], which is a standard have gotten a lot of attention recently as next generation of foundation for e-Learning, from viewpoint of AI (Chap. 2), a way to connect knowledge level and symbol level (Chap. 3), and an SCORM2004 conformed ontology-aware authoring tool (Chap. 4).

2 SCORM2004 as a Symbol Level for Learning Contents

2.1 Current State of Designing Learning Contents Conformed to SCORM2004

Currently, a typical learning content conformed to SCORM2004 has a tree structure reflected textbook structure. In such a content adaptive control is available by rules put on nodes representing chapter, section and so on.

Figure 1 shows an example of typical structure of SCORM2004-conformed learning contents. This content starts from "Pretest of brief of AI". If a learner passes the pretest, he/she will learn "the detail of AI". If not he/she will learn "brief of AI" before learning "the detail of AI". This control is implemented by the sequencing rule (preConditionRule1).

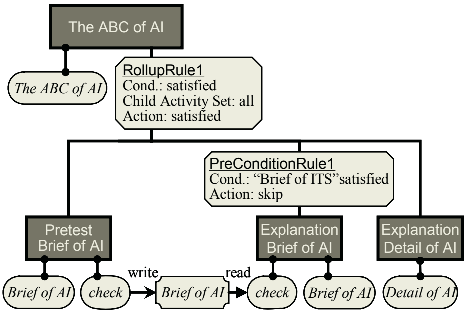


Fig. 1. An example of typical structure of SCORM2004-conformed learning contents

When an activity is finished, tracking data in the activity is aggregated to its parent activity. For instance, "The ABC of AI" aggregates tracked data from all of "Pretest Brief of AI", "Explanation Brief of AI" and "Explanation Detail of AI" (That is because a parameter of RollupRule1 Child Activity Set is set "all"). So learning result of the entire content in recorded in "The ABC of AI".

These rules in SCORM2004 realize adaptive preorder page-turner structure easily. However, the focus of IES is not control that shows all of the contents or a part of them according the preordered structure but decision making of the next activity according to a learner's status.

2.2 Lessons Learned from IES studies

IESs are educational support systems based on Artificial Intelligence technology [Wenger 87]. Typical thought of IESs is seen in the study of MENO-Tutor by Woolf [Woolf 84]. In MENO-Tutor knowledge of instructional control is described in an Augmented Transition Network (ATN). A node in the network indicates a teaching strategy or an action and is connected with other nodes that can be transit from itself. Depending on control rules referring to a learner model, instructional control is carried out by transition of the nodes. In this manner IESs have representation of its own structure of instructional knowledge. IES studies aim to generate learning sequences matching flexibility to each learner by sophisticating representation of knowledge and its learner model.

However there are some problems listing below.

I. Sharability and reusability are seriously low because a research oriented special purpose platforms are developed independently in each study. That has caused low productivity in practical aspect and few hoard of knowledge in research aspect.
II. Building an IES remains a costly work because of the complex knowledge representation and necessity of too much description of knowledge for a small learning content.

These problems stand in the way of research promotion and practical application of IESs. We suggest that the two following issues are important to solve the problems.

A. Organizing constructive concept of instructional control knowledge in IESs that allows IES designers to share their knowledge, that is, understanding others' description of knowledge easily and describing their own knowledge that the others can read easily.
B. Sharing IES platform to execute instructional control knowledge based on the constructive concept (ontology) in communities of researchers and practitioners.

2.3 Overlooking SCORM2004 from the Viewpoint of AIED Research

The Authors have developed an ontology-aware authoring tool called *i*Designer[Ikeda 97][Hayashi 04]. The study addresses the problems about making non-IES learning contents mentioned as I and II in the previous section with an approach A. Following up the previous study, this study aims to develop a high scalable user-friendly IES development environment with Sequence and Navigation specification in SCORM2004 as the basis of approach B. This section shows the authors' basic idea of SCORM2004 as a foundation of IES platforms.

An activity node in SCORM2004 is basically thought to represent learning experience learners have. An activity, that is to say, is "What to teach" from the view point of educational systems, and is "What to learn" from the viewpoint of learners. Each node represents "A material used in learning" (e.g. contents in described in a chapter, a section and a page). An activity tree represents "Structure of materials". On the other hand, in many cases, decision-making structures of IESs are "Which teaching action is better" from the system's view and "Which learning action is better" from learners'. A node represents an action and a structure represents decision-making of teaching actions, for example,. in ATN of MENO-TUTOR, "introduce", "tutor", "hack" and "complete" (Of course, if you embed "action" within "what to teach" in SCORM, it might looks like "Teaching action" in IES. But such an embedding must not be valid because it hides knowledge to select actions. This issue will be mentioned in chapter 3.)

If one wants to make a learning content to be highly adaptive to an individual learner, one must organize learning experiences with a central focus on knowledge-based decision-making structure of learning action. In this case, it is not so easy to reflect the structure to on an activity tree but not impossible. The solution is to find a way to convert selection of learning action into selection of activities in SCORM2004 and a way to convert selection structure of actions into activity tree.

This is matched with “Knowledge level and symbol level” that Newell proposed as the principle of artificial intelligent systems. Framework of description of IES knowledge is the source of intelligent behaviour of IESs and is equivalent to knowledge level. A platform that behaves intelligently based on the knowledge is equivalent to the symbol level.

The authors consider that studies of intelligent educational systems will be developed and turned into actual utilization if it is possible to build framework of knowledge level description of educational control knowledge based on SCORM2004 platform as high-scalable symbol level.

3 Building a Bridge Between a Knowledge Level and a Symbol Level

3.1 An Activity Tree as a Decision-Making Tree for Delivering Learning Object

Figure 2 indicates an example of decision-making model in knowledge level. The learning process described in the model is composed of a flow of learning. “Review”, the last part of the flow, will be done if a learner does not pass the exercise. The tree structure of “Review” represent that it is achieved either two types of alternative strategy.

As stated above, decision-making model describes a structure strategic decision-making of teaching action. Fig. 1 indicates a structure to select contents but fig. 2 indicates structure to select what to do. Clarifying knowledge to select action is the basis of IES and significant in the following two point.

A) The system can construct teaching sequences that fit for each learner’s understanding status, and
B) The system can assume a learner’s learning property through analysis of teaching action accepted by the learner.

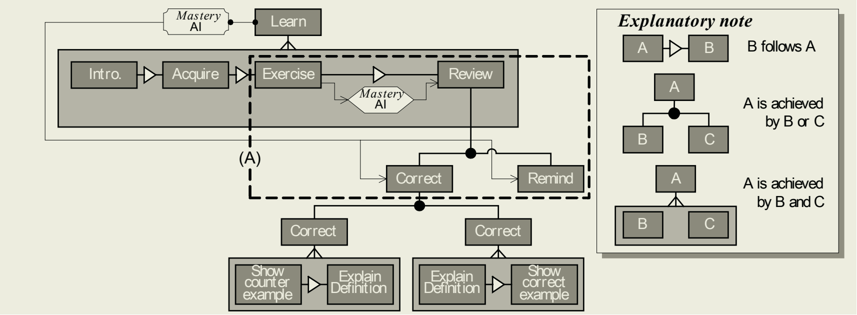


Fig. 2. An example of decision-making model in knowledge level

3.2 Symbol Level Model

Fig 3 shows the symbol level model converted from the knowledge level model shown in fig 2.

Decision-making structures are converted to structure of activity tree and sequencing rules. In fig 3 "*Exercise* AI" and "*Review* AI" are described as child activity of "Learn AI". The condition that "*Review* AI" is made available is described as PreConditionRule1.

The information collection structures are converted to roll-up rules. "*Learn* AI" must aggregate information of all of child activity so the roll-up rule1 has value "all" in child activity set.

3.3 Patterns for Designing an Activity Tree

In this section, we will organize decision-making structure and information collection structure and relate the patterns to SCORM2004 specification. This study proposes three kinds of basic decision-making and information collection patterns listing below.

(a) Alternative pattern: pattern for selecting only one activity,
(b) Process divide pattern: pattern for dividing a process, and
(c) Content divide pattern: pattern for dividing a contents

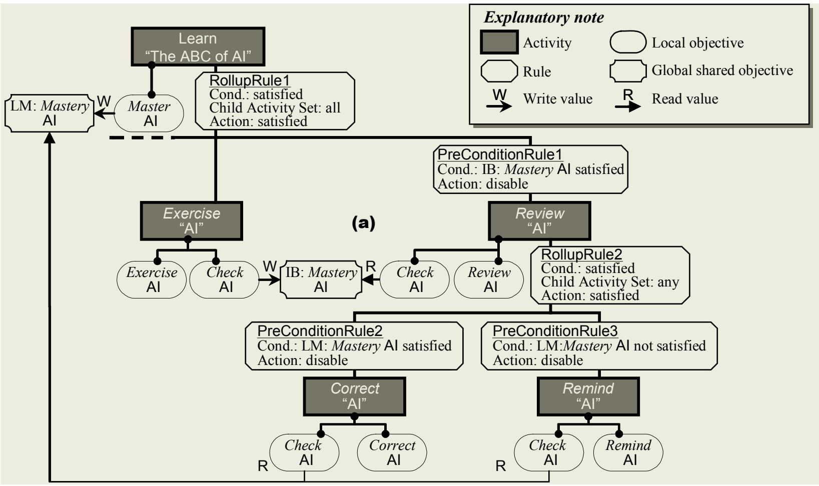
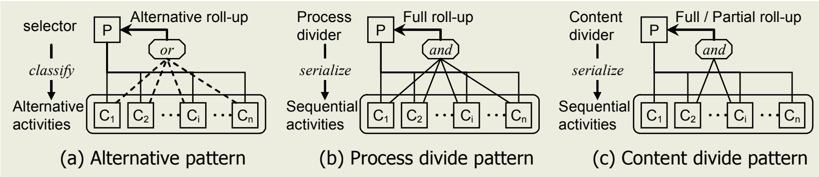


Fig. 3. Symbol level model



(a) Alternative pattern (b) Process divide pattern (c) Content divide pattern

Fig. 4. Patterns for designing an activity tree

These patterns are shown in fig.4. Each activity in the patterns are put into shape by the IES ontology (partially shown in fig. 5) which is developed our preliminary work in [Mizoguchi00][Hayashi 04]. Combination of these patterns allows designers to construct a flexible decision making model for variety of learning contents.

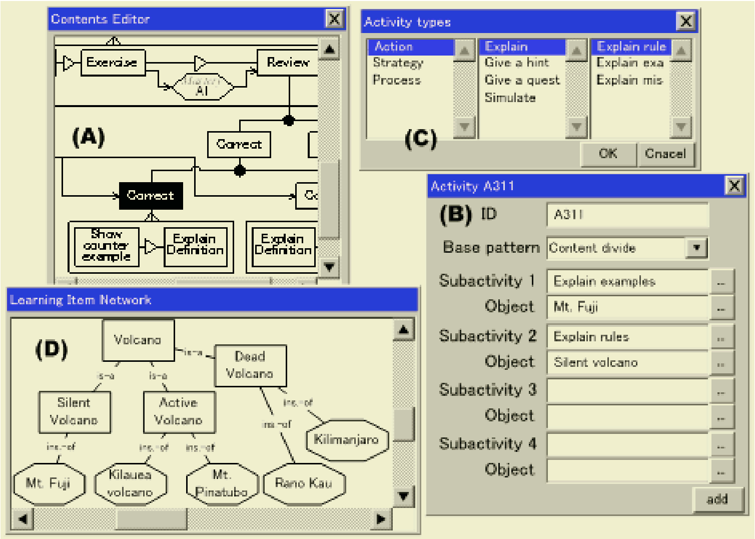
4 Toward a Knowledge Level Authoring Support

In this study we have been developing an IES authoring tool conformed SCORM2004 based on iDesigner [Hayashi 04]. Characteristics of the authoring tool are that it has not only ontology awareness [Ikeda 99] but also standard awareness for high scalability. The tool can convert author's design intention in knowledge level to the implementation in symbol level based on an ontology for learning contents, the patterns shown in fig. 4, and SCORM2004 specification.

Fig 6 indicates an image of the authoring tool. The main interface is the content editor (fig 6(A)). This shows a decision-making structure. Values of each node are set on window (B). While setting the values, authors can refer to items to be selected with windows (C) and (D).

```
■Activity
  ■Introduce
    ■Take interest
    ■Pay attention
    ■…
  ■Remind
    ■Remind concept
    ■Remind example
    ■Remind relation
    ■…
  ■Understand
    ■Remind concept
    ■Remind example
    ■Remind relation
    ■…
  ■Dismiss
  ■Correct
  ■Check
  ■Evaluate
  ■…
```

Fig. 5. Concept of activities (partial)



ID: A311
Base pattern: Content divide
Subactivity 1: Explain examples
Object: Mt. Fuji
Subactivity 2: Explain rules
Object: Silent volcano
Subactivity 3:
Object:
Subactivity 4:
Object:

Fig. 6. Interfaces of the authoring tool (imaginary)

5 Conclusion

This paper discussed learning content design with knowledge level representation on top of SCORM2004 platform as a symbol level architecture of IES decision-making structure. This approach will allow us to share and reuse academic and technical expertise in the field of AIED research on common platform. This will also contribute to develop SCORM into next generation standard specification for more adaptive and intelligent contents. Though many problems are left, for example organizing concepts related educational activities, accumulating principle or empirical knowledge of construction of activities, coordination between knowledge level and symbol level and so on, ontological engineering approach must be of assistance to do them.

References

[ADL 04] ADLNet.: *Sharable Content Object Reference Model: SCORM2004 2nd Ed.*, http:/www.adlnet.org/, 2004.

[Hayashi 04] Hayashi, Y., Ikeda, M., and Mizoguchi, R.: "A Design Environment to Articulate Design Intention of Learning Contents", International Journal of Continuing Engineering Education and Life Long Learning, Vol. 14, No. 3, pp.276-296, 2004.

[Ikeda 97] Ikeda, M., Seta, K., Mizoguchi, R.: "Task Ontology Makes It Easier To Use Authoring Tools", Proc. of Intl. Joint Conf. on Artificial Intelligence (IJCAI97), pp.23-29, 1997

[Ikeda 99] Ikeda, M., Hayashi, Y., Jin, L., Chen, W., Bourdeau, J., Seta K., and Mizoguchi R.: "Ontology more than a Shared Vocabulary", *Proc. of AIED99 Workshop on Ontologies for Intelligent Educational Systems*, pp. 1-10, 1999.

[Murray 98] Murray, T.: "Authoring Knowledge Based Tutors: Tools for Content, Instructional Strategy, Student Model, and Interface Design", *J. of the Learning Sciences*, Vol. 7, No.1, pp. 5-64. 1998.

[Murray 99] Murray, T.: "Authoring Intelligent Tutoring Systems: Analysis of the state of the art", *Int. J. of AI and Education*, Vol. 10, No. 1, pp. 98-129, 1999.

[Murray 03] Murray, T., Blessing, S., and Ainsworth, S. (eds.): *Authoring Tools for Advanced Technology Learning Environments*, Kluwer Academic Publishers, Netherlands, pp.439-466, 2003.

[Mizoguchi 00] Mizoguchi, R. and Bourdeau, J.: "Using Ontological Engineering to Overcome AI-ED Problems", *Int. J. of Artificial Intelligence in Education*, Vol.11, No.2, pp.107-121, 2000.

[Newell 82] Newell, A.: The Knowledge Level, *Artificial Intelligence*, Vol. 18, pp. 87-27, 1982.

[Wenger 87] Wenger, E.: *Artificial intelligence and tutoring systems: Conceptual and cognitive approaches to the communication of knowledge*, Morgan Kaufmann Publishers, 1987.

[Woolf 84] Woolf, B. P, McDonarld, D. D.: "Building a computer tutor: design issues", IEEE Computer, Vol. 17, no 9, pp 61-73, 1984.

Analyzing Domain Expertise by Considering Variants of Knowledge in Multiple Time Scales

Jun-Ming Chen[1], Gwo-Haur Hwang[2], Gwo-Jen Hwang[3], and Carol H.C. Chu[1]

[1] Information Management Department, National Chi Nan University
Puli, Nantou, Taiwan 545, R.O.C.
{s2213519,carol}@ncnu.edu.tw
[2] Information Management Department, Ling Tung College
Taichung, Taiwan 408, R.O.C
ghhwang@mail.ltc.edu.tw
[3] Department of Information and Learning Technology, National University of Tainan
Tainan, Taiwan 700, R.O.C.
gjhwang@mail.nutn.edu.tw

Abstract. Knowledge acquisition is known to be a critical bottleneck in building expert systems. In past decades, various methods and systems have been proposed to efficiently elicit expertise from domain experts. However, in building a medical expert system, disease symptoms are usually treated as time-irrelevant attributes, such that much important information is abandoned and hence the performance of the constructed expert systems is significantly affected. To cope with this problem, in this paper, we propose a time scale-oriented approach to eliciting medical knowledge from domain experts. The novel approach takes the time scale into consideration, such that the variant of disease symptoms in different time scales can be precisely expressed. An application to the development of a medical expert system has depicted the superiority of our approach.

1 Introduction

Expert systems have been applied to many problem-solving activities such as decision making, designing, planning, monitoring, diagnosing, and training activities. Subject domains that are supported by experts systems include bioengineering, defense, education, engineering, finance, and medical diagnosis. Among those domains, medical diagnosis is one of the most popular and successful applications of expert systems. In 1973 , the MYCIN project was conducted in Stanford University, which has become a well-known medical expert system for diagnosing infectious diseases (Buchanan & Shortliffe, 1985). The success of MYCIN project has encouraged the advent of medical expert systems. In the past decades, a number of expert systems have been developed to cope with medical diagnosis problems; for example, FRBS-GP is a fuzzy rule-based system for diagnosing aphasia's subtypes and the classification of pap-smear examinations (Jantzen et al., 2002); TheraTrac 2 is a system for microbiological data validation and real-time emission of alarms (Theratrac, 2001); Vitek is an expert system integrated into particular analytical instruments for test result validation (Vitek, 2001); Neural Ensemble-based Detection (NED) is used

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1324–1330, 2005.

to identify lung cancer cells in the images of the specimens of needle biopsies (Zhou et al., 2002); DNSev is an expert system that helps laboratory physicians in validating biochemical analysis results (Storari et al., 2003).

In building an expert system, the critical bottleneck is to obtain the knowledge of the special domain from the domain experts, which is called knowledge acquisition. Although various knowledge acquisition methods have been proposed in the past decades, in building medical expert systems, disease symptoms are usually treated as time-irrelevant attributes, such that much important information is abandoned and hence the performance of the constructed expert systems is significantly affected. To cope with this problem, we shall propose a time scale-oriented approach to eliciting medical knowledge from domain experts. The novel approach takes the time scale into consideration, such that the variant of disease symptoms in different time scales can be precisely expressed. Experimental results show that our approach can achieve much better performance than conventional knowledge acquisition approach.

2 Relevant Researches

In the past decades, many knowledge acquisition systems were developed to build rapid prototypes and to improve the quality of the elicited knowledge, e.g., ETS (Boose, 1985), SALT (Marcus, 1987), NeoETS (Boose et al., 1986), KNACK (Klinker, 1987), AQUINAS (Boose & Bradshaw, 1987), KITTEN (Shaw & Gaines, 1987), KSSO (Gaines, 1987), ASK (Gruber, 1988), WordNet (Millar, 1990), KADS (Schreiber et al., 1993), MCRDR (Kang, 1996), KAMET (Cairó, 1997), Med-Frame/CADIAG-IV (Leitich et al., 2001), MPKBS (Cheung et al., 2003), FFAS (McIvor et al., 2004), DDosITS (Lin & Tseng, 2004), EKEL (Cao et al., 2004), MKBS (Yan et al., 2004), ANDES (Zhang et al., 2004), VIBEX (Yang et al., 2005). Most of these systems were developed based on the repertory grids method originated from Kelly's Personal construct theory (Kelly, 1955), which assists in identifying different objects in a domain and distinguishing among these objects.

A repertory grid is represented as a matrix whose columns have elements labels and whose rows have construct labels. After the set of constructs is ready, the expert is asked to fill the grid with ratings. A 5-scale rating mechanism is usually used in filling the grid; i.e., each rating is an integer ranging from 1 to 5, where "1" represents that the element is very likely to have the trait; "2" represents the element may have the trait; "3" represents "unknown" or "no relevance"; "4" represents that the element may have the opposite characteristic of the trait; "5" represents that the element is very likely to have the opposite characteristic of the trait. Table 1 shows an example of a repertory grid for gastrointestinal diseases.

Table 1. Illustrative example of a repertory grid for gastrointestinal diseases

	Appendicitis	Enteritis	
Diarrhea	2	2	Not Diarrhea
Lower abdominal pain	3	5	Not Lower abdominal pain
Abdominal distension	4	4	Not Abdominal distension
Fever	5	4	Not Fever

3 Multi-dimensional Repertory Grid Approach

One critical problem in applying the existing knowledge acquisition methods to elicit medical expertise is the treatment of those diseases as time-irrelevant elements. In the real world, the features of diseases are likely to change from time to time. For example, in the earlier stage of having appendicitis, the symptoms of the patients might be diarrhea, lower abdominal pain, nausea, vomit, abdominal distension, stomache and fever. Several days later, the symptoms become Teulerness R.L.Q. of abdominal, Teulerness R.L.Q to return abdominal, Leukocytosis, etc. One or two days later, these uncomfortable symptoms might gradually ease off. Sometimes the second stage of a disease might looks similar to the first stage of other diseases; therefore, it would be improper to ask a medical expert to describe the disease symptoms without considering the time scale.

To precisely elicit time-variant medical diagnosis knowledge from domain experts, a Multi-Dimension Repertory Grid (MDRG) is proposed in this section, which takes time scale as a new dimension in the extended repertory grid. In addition to time scale, MDRG takes importance degree for each construct to each element in different time scales into consideration, such that more embedded knowledge can be explicitly presented.

Let e_i^t denote t-th stage period of element (or disease) e_i and c_j denote a construct (or symptom), where i = 1 to n, and j = 1 to m. Each MDRG entry is a triplet that consists of three values: a rating to indicate the relevance of e_i^t and c_j, a certainty degree for giving the rating and an impact factor to represent the importance of c_j to e_i^t, which are represented by the following three functions:

Let e_i^t denote t-th stage period of element (or disease) e_i and c_j denote a construct (or symptom), where i = 1 to n, and j = 1 to m. Each MDRG entry is a triplet that consists of three values: a rating to indicate the relevance of e_i^t and c_j, a certainty degree for giving the rating and an impact factor to represent the importance of c_j to e_i^t, which are represented by the following three functions:

(1) Rating (e_i^t, c_j): the degree of relevance for disease $\mathrm{e_i}$ in t-th time scale to symptom $\mathrm{c_j}$, ranging from 1 to 5: “1” represents that the element is very likely to have the opposite characteristic of the trait; “2” represents the element may have the opposite characteristic of the trait; “3” represents “unknown” or “no relevance”; “4” represents that the element may have the trait; “5” represents that the element is very likely to have the trait.

(2) Certainty (e_i^t, c_j): the degree of certainty for giving the rating Rating (e_i^t, $\mathrm{c_j}$), which is either ‘S’ or ‘N’ representing ‘sure’ or ‘not sure’.

(3) Impact_ factor (e_i^t, c_j): the degree of importance for symptom $\mathrm{c_j}$ to disease $\mathrm{e_i}$ in t-th time scale. Impact_ factor (e_i^t, $\mathrm{c_j}$) can be one of the following values: ‘X’ represents no relationship between the disease and the symptom; ‘D’ means that the symptom dominates the disease, i.e., if the value of the symptom is not matched, it is impossible for the disease to be implied; an integer, ranging from 1 to 5, indicates that the symptom is of some degree of importance to the disease, but does not dominate the disease.

Table 2. Illustrative example of a MDRG

Disease	Appendicitis		Enteritis
Stage	T_1	T_2	T_1
Diarrhea	2,S,3	1,S,X	4,S,5
Lower abdominal pain	5,S,D	2,N,3	5,S,1
Abdominal distension	4,S,2	1,S,X	4,S,4
Fever	5,S,4	4,S,3	4,N,2

An example of a MDRG is given in Table 2, where Appendicitis in the *k*-th time scale is recorded as T_k of Appendicitis or Appendicitis ^{k}. For example, the second time scale of Influenza is recorded as T_2 of Influenza or Influenza^2.

It can be seen that the disease in different time scales may have difference symptoms. For example, Rating (Appendicitis^1, Abdominal distension) = 4 while Rating (Appendicitis^2, Abdominal distension) = 1 implies that the symptom "Abdominal distension" is apparent in the first time scale but is not apparent in the second time scale of Influenza.

Moreover, Certainty (Enteritis^1, Fever) = "N" indicates that the domain expert is not very certain while giving Rating (Enteritis^1, Fever) = 4; Certainty (Enteritis^1, Diarrhea) = "S" indicates that the domain expert is very certain while giving Rating (Enteritis^1, Diarrhea) = 4.

In addition, Impact_ factor (Appendicitis^1, Lower abdominal pain) = "D" indicates that the symptom "Lower abdominal pain" dominates the disease "Appendicitis in the first time scale"; that is, if the symptom "Lower abdominal pain" is not apparent (since Rating (Appendicitis^1, Lower abdominal pain) = 5), it is impossible for the disease "Appendicitis in the first time scale" to be implied. Impact_ factor (Appendicitis^1, fever) = 4 > Impact_ factor (Appendicitis^1, Abdominal distension) = 2 implies that the symptom "fever" is more important than the symptom "Abdominal distension" to the disease "Appendicitis in the first time scale"; therefore, to conclude "the disease is Appendicitis in the first time scale", the negation of the fact "the patient has fever" will be affected the degree of certainty more than that the negation of the fact "patient has Abdominal distension".

4 Experiments and Evaluations

To evaluate the performance of our novel approach, two knowledge bases constructed by employing traditional repertory grid approach and the new approach, respectively. The application domain is "the diagnosis of gastrointestinal diseases", which includes thirteen diseases: 1- Gastroenteritis, 2- Gastritis, 3- Gastric ulcer, 4- Gastroesophaged Reflux disease: GERD, 5- Maldiggestion, 6- Malabsorption, 7- Duodenal ulcer, 8- Appendicitis, 9- Lower Drago and Montoneri, 10- Irritable colon, 11- Ulcerative Colitis, 12- Hiatus hernia, and 13- Amebic dysentery. Nineteen symptoms are used to identify those diseases, including "diarrhea", "flatulence", "lower abdominal pain", "nausea", "vomit", "abdominal distension", "stomache", "fever", "Teulerness R.L.Q. of abdominal", "Teulerness R.L.Q to return abdominal", "abdominal pain", "aqueous stool", "mucinous stool", "stool is blood-stained", "leukocytosis", "air belching", "melena", "bloody stool" and "acid regurgitation".

Thirty-three cases given by the medical expert were used to test the performance of the constructed knowledge bases. Table 8 shows the diagnosis results given by the medical expert and the expert systems constructed by employing the repertory grid approach and the time scale-oriented approach, where an ‘X” indicates that the expert system can not reach any conclusion, an integer number represent the disease number, and the “D-T” format represents the disease numbered D in T-the time scale. For example, “1-2” means “Gastroenteritis in 2nd time scale”.

From the data given in Table 3, it can be seen that the expert system constructed with the time scale-oriented approach not only can correctly diagnose the disease, but can also indicate the time scale of the disease. Nevertheless, the expert system with conventional repertory grid approach has provided several answers that are apparently different from those given by the medical expert.

Table 3. Testing results of the old and new prototypes

Case number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Human expert	1	1	2	2	3	3	4	4	5	5	6	6	6	7	7	8	8
Repertory grid	1	10	2	X	3	7	4	X	1	5	1	6	5	4	7	1	8
MDRG	1-2	1-1	2-2	2-2	3-2	3-3	4-2	4-3	5-1	5-2	6-2	6-1	6-2	7-2	7-1	8-1	8-2

Case number	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Human expert	8	8	9	9	10	10	10	11	11	12	12	12	12	13	13	13
Repertory grid	X	8	9	10	11	10	9	13	11	12	12	X	12	X	9	13
MDRG	8-1	8-1	9-2	9-3	10-1	10-1	10-2	11-3	11-2	12-1	12-2	12-1	12-1	13-1	13-2	13-2

Table 4 shows comparison for the correct-diagnosis rates of the two approaches. It can be seen that the performance of the time scale-oriented approach is significantly better than the traditional knowledge acquisition approach.

Table 4. Correct-diagnosis rate of each approach

	Correct-diagnosis rate
Repertory grid approach	51.5%
Time scale-oriented approach	100%

6 Conclusion

To accurately eliciting medical diagnosis knowledge from domain experts, we propose a time scale-oriented approach, MDRG, in this paper. The novel approach attempts to take time scale into consideration while eliciting knowledge from medical experts, such that the variant of disease symptoms in different stages can be precisely expressed. In order to evaluate the performance of our approach, an experiment on a practical application was conducted. Based on the experimental results, it can be seen that MDRG can significantly improve the quality of the knowledge base. Currently, we are trying to develop and extend our approach to accumulate more clinical experiences in developing medical expert systems.

Acknowledgement

The authors are very grateful to Dr. J.C. Huang for providing precious knowledge and conducting the experiment of this study. Part of this study is supported by the National Science Council of Taiwan, R.O.C., under contract numbers NSC-93-2524-S-260-001 and NSC-93-2524-S-009-004-EC3.

References

1. Boose, J.H.: A knowledge acquisition program for expert systems based on personal construct psychology. International Journal of Man Machine Studies, 23 (1985)
2. Boose, J.H. and Bradshaw, J.M.: NeoETS: Capturing expert system knowledge in hierarchical rating grids. IEEE Expert System in Government Symposium (1986)
3. Boose, J.H. and Bradshaw, J.M.: Expertise transfer and complex problems: using AQUINAS as a knowledge-acquisition workbench for knowledge-based systems. International Journal of Man Machine Studies, Vol. 26. (1987) 3–27
4. Buchanan, B. and Shortliffe, E. (eds.): Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Paper. Reading, MA: Addison-Wesley (1985)
5. Cairó, O.: The KAMET methodology: A modeling approach for knowledge acquisition. EXPER SYS-97, Subject Editor: Smith., P. Technology Transfer Series, Series Editor: Niku-Lari, A., IITT international (1997)
6. Cao, C., Wang, H. and Sui, Y.: Knowledge modeling and acquisition of traditional Chinese herbal drugs and formulae from text. Artificial Intelligence in Medicine, Vol. 32. No. 1. Sep. (2004) 3-13
7. Cheung, C.F., Lee, W.B., Wang, W.M., Chu, K.F. and To, S.: A multi-perspective knowledge-based system for customer service management. Expert Systems with Applications, Vol. 24. No. 4. May. (2003) 457-470
8. Gaines, B.R.: An overview of knowledge-acquisition and transfer. International Journal of Man Machine Studies, Vol. 26. (1987) 453–472
9. Gruber, S.T.: The Acquisition of Strategic Knowledge. Academic Press Inc (1988)
10. Jantzen, J., Axer, H., von Keyserlingk, D.G.: Diagnosis of aphasia using neural and fuzzy techniques. In: Zimmermann, H-J, Tselentis, G., van Someren, M., Dounias, G. (eds.): Advances in computational intelligence and learning. Massachussets: Kluwer Academic Publishers (2002) 461–74
11. Kang, B.: Multiple classification ripple down rules. PhD Thesis, University of New South Wales (1996)
12. Kelly, G.A.: The psychology of personal constructs, Vol. 1. NY: W.W Norton (1955)
13. Klinker, G., Bentolila, J., Genetet, S., Grimes, M. and McDermott, J.: KNACK – Report-driven knowledge acquisition. International Journal of Man Machine Studies, Vol. 26. (1987) 65–79
14. Lee, R.C.T., Chang, R.C. and Tseng, S.S.: Introduction to the Design and Analysis of Algorithms. Chapter 7. Reading, MA: Prentice Hall (1989)
15. Lin, S-C., Tseng, S-S.: Constructing detection knowledge for DDoS intrusion tolerance. Expert Systems with Applications, Vol. 27. No. 3. Oct. (2004) 379-390
16. Leitich, H., Kiener, H.P., Kolarz, G., Schuh, C., Graninger, W., Adlassnig, K-P.: A prospective evaluation of the medical consultation system CADIAG-II/RHEUMA in a rheumatological outpatient clinic. Methods Inform Med, 40 (2001) 213–20
17. Marcus, S.: Taking backtracking with a grain of SALT. International Journal of Man-Machine Studies, 26 (1987) 383–398

18. McIvor, R.T., McCloskey, A.G., Humphreys, P.K. and Maguire, L.P.: Using a fuzzy approach to support financial analysis in the corporate acquisition process. Expert Systems with Applications, Vol. 27. No. 4. Jul. (2004) 533–547
19. Millar, A.: WordNet: an on-line lexical resource. Journal of Lexicography, 3(4) (1990)
20. Schreiber, G., Wielinga, B. and Breuker, J.: KADS: A principle approach to knowledge-based system development. London: Academic Press (1993)
21. Shaw, M.L.G. and Gaines, B.R.: KITTEN: Knowledge initiation and transfer tools for experts and novices. International Journal of Man-Machine Studies, 27 (1987) 251–280.
22. Storari, S., Lamma, E., Mancini, R., Mello, P., Motta, R, Patrono, D. and Canova, G.: Validation of biochemical laboratory results using the DNSev expert system. Expert Systems with Applications, Vol. 25. No. 4. Nov. (2003) 503-515
23. Theratrac, Biomerieux. www.theratrac.com, Accessed 10 July 2001.
24. Vitek, Biomerieux. www.biomerieux.com, Accessed 10 July 2001.
25. Yan, H., Jiang, Y.T., Zheng, J., Fu, B.M, Shao, S. and Peng, C.: The internet-based knowledge acquisition and management method to construct large-scale distributed medical expert systems. Computer Methods and Programs in Biomedicine, Vol. 74. No. 1. Apr. (2004) 1-10
26. Yang, B-S., Lim, D-S. and Tan, A.C.C.: VIBEX: an expert system for vibration fault diagnosis of rotating machinery using decision tree and decision table. Expert Systems with Applications, Vol. 28. No. 4. May. (2005) 735-742
27. Zhang, D., Lee, V., Friedel, J. and Keyser, R.: Automated Facts Generation From Raw Data: A Perspective From The ANDES Project. Electronic Notes in Theoretical Computer Science, Vol. 25. No. 1. Feb. (2004) 1-10
28. Zhou, Z-H., Jiang, Y., Yang, Y-B., Chen, S-F.: Lung cancer cell identification based on artificial neural network ensembles. Artificial Intelligence in Medicine, Vol. 24. No. 1. Jan (2002) 25–36

A New Algorithm to Discover Page-Action Rules on Web

Heng-Li Yang and Qing-Fung Lin

64, Sec. 2, Chihnan Rd., Mucha Dist, Taipei 116, Taiwan
{yanh,qflin}@nccu.edu.tw

Abstract. In the past, a number of researches have been devoted to web mining. However, only page browsing behaviors have been studied. The actions performed in the pages were omitted. To the best of our knowledge, no previous researches were able to find out page-action browsing sequences. This paper proposes an algorithm, called WebPAN, to analyze customers' browsing pages and their action paths. The algorithm's efficiency was examined in our prototype. This study would help website managers to restructure their website layouts or advertisement positions more correctly in electronic commerce.

1 Introduction

With the rapid development of Internet, the data mining on WWW, called Web mining, has obtained a great number of attentions. There are three types of Web mining: Web content mining on page contents or search results, Web structure mining on Website reference hyperlinks, and Web usage mining on Web logs [4]. Our study belongs to the third type. However, in the past, researches (e.g., [2, 3, 5, 6, 8, 9]) have paid most attentions to browsing page sequences. To the best of our knowledge, no previous researches focused on the performed actions of the browsed pages.

Take a Web log in Table 1 as an example. Those in the brackets of the third column are the performed actions on the associated pages. Traditionally, focusing only on the browsed pages, we can only find a rule such as "If an user browses *On Sale* page and enter *Product Information* page, there are 66% that he (she) would enter *View Basket* page." However, examining the details and considering both browsed pages and performed actions, we would find a rule such as "If an user views price on the *On Sale* page and checks *payment ways* on the *Product Information* page, there are 100% that he (she) would submit order on the *View Basket* page". This study aims to find those path-action rules that would help Website managers to understand users' behaviors more clearly.

2 The Proposed WebPAN Algorithm

Our algorithm is modified from Apriori [1] to find the association rules. The original Apriori algorithm first computes the support value of Candidate Sequences (*CS*), determines any support greater than large threshold to produce Large Sequence (*LS*), then produces CS_{K+1} from LS_K, iterates until no more *LS*. Before describing our WebPAN algorithm, we will first introduce some concepts as follows.

PA_Sequence Definition. An user could perform a number of actions in a Web page. Suppose the logs of Web users' browsed pages and performed actions are kept. An action is not limited to confirmed transaction action [8]. It could be like "clicking a

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1331–1337, 2005.

button", "submitting a request", "choosing an item from a menu", "moving a cursor to certain area", etc., but not eye actions. Additional self-developed program is needed to capture these records. After pre-processing, each user transaction record could be transformed into a form such as <*A(bc)B(ac)*>, which means the user performed actions *b, c* in sequence on page *A*, then performed actions *a, c* in sequence on page *B*. A *PA_Sequence* contains only Page_ID and Action_ID. There is no sequence such as <*A(bc)A(ca)*>. Instead, we would transform it into <*A(bcca)*>. In addition, *s*uppose *S* is the set of all *PA_Sequence*s formed by Page_IDs and Action_IDs. Here are our formation rules:

Rule 1 For any s∈ *S*, *s* would not contain any action without its related page.

Rule 2 For any s∈ *S*, *s* would not contain any page without its performed action.

Table 1. An example of browsed pages and performed actions

SessionId	Page Path	Page-Action Path
A00001	*On Sale → Product Information*	*On Sale [*View Price*] →Product Information[*View Detail Information*]*
A00002	*On Sale → View Basket*	*On Sale[*View Price → Add to Basket*] → View Basket[*Enter Order Information→ Submit Order*]*
A00003	*On Sale → Product Information → View Basket*	*On Sale[*View Price*] → Product Information[*View Detail Information → View Payment Ways*] → View Basket[*Enter Order Information → Submit Order*]*
A00004	*Product Information → View Basket*	*Product Information[*View Detail Information*] → View Basket[*Enter Order Information → Submit Order*]*
A00005	*On Sale → Product Information→View Basket*	*On Sale[*View Price*] → Product Information[* View Payment Ways*]→View Basket[*Enter Order Information → Submit Order*]*

PA_Sequence Length. A *PA_Sequence* has two dimensions of page and action. Considering only the browsed page sequence, a *PA_Sequence* would be reduced to *PAGE(s)*, which was the traditional research focus. On the other hand, regardless of the related pages, a *PA_Sequence* would be also reduced to *ACTION(s)*. We can compute both lengths. For example, to the sequence <*A(bc)B(ac)C(ba)*>, the former *PAGE(s)=ABC* has a length 3; the latter *ACTION(s)=bcacba* has a length 6; the total length of the *PA_Sequence* is 9.

PA_Pattern. A *PA Element* is comprised of one *Page_ID* and one *Action_ID*. A *PA Pattern* is comprised of one or several *PA Elements*. A *PA Pattern* could be part of one target, a CS or LS, in our algorithm. For example, *[AaBb]* is a pattern, which means that if a user performs an action *a* on page *A*, then sequentially performs an action *b* on page *B*.

2.1 PA Scan

In general, scanning a sequence for fitting a pattern can consider three aspects: happening (whether those elements happen?), order (whether those elements happen in order), adjacency (whether no other stuffs happen between the target elements). In this paper, we would allow actions to be performed non-successively, but pages to be browsed successively. Take the table 2 as an example. Only *S001* satisfies pattern *[AcBc]*. *S004* does not because page *C* was browsed between *A* and *B*. *S001*, *S003*, and *S004* satisfy pattern *[AcAb]*.

Table 2. An example of PA_Sequence (capitalized letters are pages, small letters are actions)

Session_Id	PA_Sequence
S001	A(acb)B(ac)A(bc)
S002	A(a)B(cab)C(ac)
S003	B(ca)C(abc)A(cb)B(ab)
S004	A(cb)C(acb)B(ca)
S005	A(a) B(cb)

Another issue is how to compute support values. If a sequence has pattern occurrences *N* times, whether its support value becomes *N* or just 1? Considering the happening weights, our algorithm would count *N*. For example, the support value Pattern *[Aa]* in a sequence *<A(aba)B(ac)A(ac)>* would be 2 because *Aa* appears twice. Note that this paper does not consider the action happening weights on the same page.

The core of our WebPAN algorithm is PA_Scan_N (in Java-like codes as follows) to determine the support value of a *Seq* for fitting *Pat*. Varible *i, j* are page cursors of *Seq* and *Pat*, respectively; variable *k* is the action cursors of a pattern; variable *x* is the scanned cursors in a certain page; variable *key* is used to determine whether part of *Seq* satisfies the *Pat.* There are three loops: (1) L02-L15 checks all possible pages in the *Seq*; (2)L05-L12 deals with page path in the *Pat*; (3)L08-L11 compares actions in certain page. L02 compares *N* times, where *N*=the page length of *Seq*'s minus the page length of *Pat*'s then plus 1. Assume actions on (i+j)'s page of *Seq* are *XS*, L09 would examine *XS* beginning the *x*th's position, and return the first position that the compared action *Xpat[k]* appears (returning -1 if fails).

```
PA_Scan_N( Pat as Pattern, Seq as Sequence) as Integer{
00  int i,j,k,Support = 0;
01  int PPLen = Pat.Page.length;
02  For (i=0 ;i<=(Seq.Page.length-PPLen);i++) {
03    If(Pat.Page==Seq.Page.substring(i,PPLen)){
04    int key=1
05    For (j=0;j<=PPLen-1;j++) {
06      int x=0;
07      String XPat=Pat.Action.atPage[j];
08      For (k=0;k<=XPat.length-1;k++){
09        x= Seq.Action.atPage[i+j].indexof(Xpat[k],x);
10        If(x==-1){key=0; j=PPLen; k=XPat.length}
11     } }//End For Loop L08~L11 and For Loop L05~L11
12  If (key == 1) { Support++}
13  } }//End If L03~L13 and For Loop L02~L13
14 Return Support;
15 }//End Function
```

2.2 Large Threshold

Traditionally, while looking for possible association rules, the threshold could be set as multiplying certain percentages by number of total records. However, while an algorithm considers possibility of *N* occurrences, the support of one record might be large than 1, its threshold is not a simple value. In iteration, each pass has different threshold. However, our case is more complicated because in the same pass possible candidate sequences might have different page lengths and so different thresholds. In addition, our algorithm adopts the idea "next pass large threshold" (looking one pass ahead) [2] to identify "next pass large sequence".

2.3 PA Join

The join to produce candidate sequence in our algorithm is modified from Srikant and Agrawal [7], that is, LS_K joins LS_K to produce CS_{K+1}. Those LS_K are *parent patterns*, CS_{K+1} are *child pattern*. The *child pattern* must cover the semantics of *parent patterns*. However, because our patterns include pages and actions, if two patterns can be joined, they must be concatenable (as figure 1) or mergable (as figure 2). It can also be proven that a pattern $CS_{i \text{ where } i>=3}$ must be produced either by a pair of concatenable or mergable patterns in its former CS_{i-1} pass.

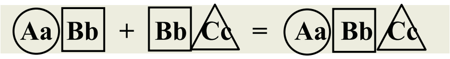


Fig. 1. A pair of concatenable patterns

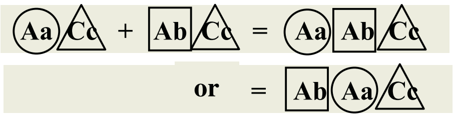


Fig. 2. A pair of mergable patterns

3 An Example

Assume a database has a Web log as table 2. Also, assume the percentage threshold is 25%. We want to apply WebPAN algorithm to perform Web mining. The WebPAN algorithm would call PA_Scan_N to produce *LS0*, *CS1*, *LS1*, *CS2*, *LS2* in sequence, and call PA_Join to produce *CS(n)* for *n*>=3.

As in Table 3, to obtain threshold vales, we must first compute total possible supports (TPS), i.e., number of possible combinations for different page lengths. Then, we apply PA_Scan_N to obtain candidate sequences CS_1 as table 4. In the third column of table 4, there are two thresholds: this pass and next pass. By applying the approach of "looking one pass ahead", some candidates, e.g., *Bb*, which are not large in this pass *i*, but may be large in next one pass, would be retained in LS_i and marked as "next 1 large". Those "next 1 large" patterns cannot be used to generate rules, but would be used to produce next pass *CS*. Therefore, LS_1 include the following sequences: *Ab*, *Ac*, *Ba*, *Bc* (large), and *Aa*, *Bb*, *Ca*, *Cc* (next 1 large). Those in LS_1 would be joined to produce CS_2, such as, *AaBc* (large), *AaBb*, *BcCa*, etc. (next 1 large). And then, we can get CS_3, such as, *AaBcBb*, *AaBbBc*, *AaBcCa*,etc. The algorithm will stop at CS_4 and get rules, such as "if an user performs *c* on page *B*, there are 60% chances that he (she) came from page *A* and performed action *a* there." (i.e., *Bc*→ *AaBc*).

Table 3. Threshold Computation

Page Length	1	2	3	4
TPS	15	10	5	1
Threshold= TPS*MinS	3.75	2.5	1.25	0.25

Table 4. Candidate Sequences CS_1

CS	Page Length	This/Next Threshold	Support	State
Aa	1	3.75 / 2.5	3	Next1 Large
Ab	1	3.75 / 2.5	4	Large
Ac	1	3.75 / 2.5	4	Large
Ba	1	3.75 / 2.5	5	Large
Bb	1	3.75 / 2.5	3	Next1 Large
Bc	1	3.75 / 2.5	5	Large
Ca	1	3.75 / 2.5	3	Next1 Large
Cb	1	3.75 / 2.5	2	
Cc	1	3.75 / 2.5	3	Next1 Large

4 Performance Evaluation

To study the PA_Scan_N performance, a prototype system was implemented by Java and tested on a PC with AMD Athlon 2400+ and 256 MB memory under Window XP operation system. The run time starting with the scan was computed by the computer. We also compared two cases: "looking one pass ahead" (solid lines in the following figure) and "not looking one pass ahead" (dot line). Simulated data were generated. Several parameters were considered: number of transactions (N), minimum support (MinSup), average length of browsed pages (AVGPL), average length of performed actions (AVGAL), average variance of browsed pages (AVGPVar), and average variance of performed actions (AVGAVar).

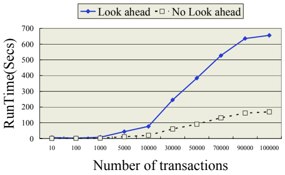


Fig. 3. Data volume effect

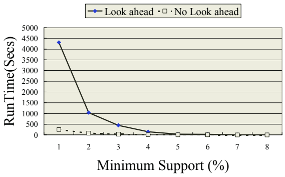


Fig. 4. Minimum support effect

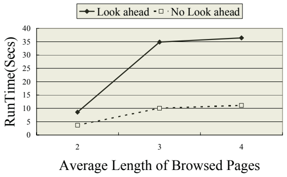


Fig. 5. Page length effect

Figure 3 was drawn to show run times for different data volumes with MinSup=25%, AVGPL=3, AVGAL=4, AVGPVar=0, and AVGAVar=0. The relationship is linear, and "looking ahead" would take more time. Figure 4 was drawn to show run times for different minimum supports with N=1000, AVGPL=3, AVGAL=4, AVGPVar=1, and AVGAVar=2. It indicates that run time would increase as the minimum support decreases. Figure 5 was drawn to show run times for different average lengths of browsed pages with MinSup=20%, AVGAL=4, and AVGAVar=0. It indicates that the longer the page path is, the more run time it takes. Figure 6 was drawn to show run times for different average lengths of performed actions with MinSup=10%, AVGPL=3, and AVGPVar=0. It also indicates that the longer the action path is, the more run time it takes. Figure 7 was drawn to show run times for different average variances of browsed pages with MinSup=15%. Figure 8 was drawn to show run times for different average variance of performed actions with MinSup=15%. Both figures indicate that neither the dispersion of path lengths nor the dispersion of action lengths would affect performance.

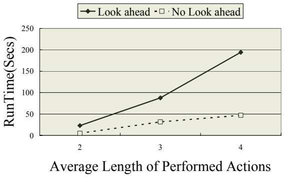


Fig. 6. Action length effect

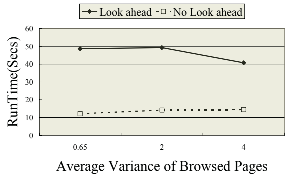


Fig. 7. Page length dispersion

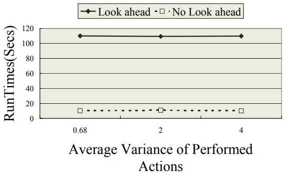


Fig. 8. Action length dispersion

5 Conclusions

This study has proposed a new algorithm WebPAN to discover path action paths on Web. We can find rules such as "suppose that an user performs an action *a* on page *A*, there are 70% chances that he (she) came from page *B* and performed an action *b* there", or "suppose that an user performs an action *d* on page *A*, there are 80% chances that he (she) would perform an action *e* on the following page *B*". Such types of rules would be helpful for designing Web layout, placing advertisements, giving promotions, etc. In the future research, we will consider the action happening weights on the same page, other rules to require actions to be performed successively, or allow pages to be browsed non-successively, etc.

References

1. Agrawal, R. Srikant, R.: Fast Algorithms for Mining Association Rules. Proc. of the 20th Int'l Conference on Very Large Databases. (1994)
2. Chen, S-S., Hsu, P-Y, Chen, Y-L.: Mining Web Traversal Rules with Sequences. MIS Review. 9 (1999) 53-71
3. Chen, M-S., Park, J-S., Yu, P. S.: Efficient Data Mining for Path Traversal Patterns. IEEE Trans. on Knowledge and Data Engineering. 10(2) (1998) 209-221
4. Cooley, R., Mobasher, B., Srivastava, J.: Web Mining: Information and Pattern Discovery on the World Wide Web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97). (1997)
5. Cooley, R., Mobasher, B., Srivastava, J.: Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns. Proceedings of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX-97*).* (1997)
6. Hsieh, C. C., Chang, C.T.: An Enhanced Transaction Identification Module on Web Usage Mining. Asia Pacific Management. (2001) 241-252
7. Srikant, R. Agrawal, R.: Mining Sequential Patterns: Generalizations and Performance Improvements. Proc. of the Fifth Int'l Conference on Extending Database Technology (EDBT). (1996)

8. Yun, C.H., Chen, M.S.: Using Pattern-join and Purchase-Combination for Mining Transaction Patterns in an Electronic Commerce Environment. The 24th Annual International Conference On Computer Software and Applications. (2000) 99-104
9. Zhang, W., Xu, B., Song, W., Yung, H., Liu, K.: Data Mining Algorithms for Web Prefetching. Proceeding of the First International Conference On Web Information Systems Engineering. 2 (2000) 34-38

Efficient Remining of Generalized Association Rules Under Multiple Minimum Support Refinement

Ming-Cheng Tseng[1], Wen-Yang Lin[2], and Rong Jeng[1]

[1] Institute of Information Management, I-Shou University, Kaohsiung 840, Taiwan
clark.tseng@msa.hinet.net, rjeng@isu.edu.tw
[2] Department of Computer Science and Information Engineering,
National University of Kaohsiung, Kaohsiung 811, Taiwan
wylin@nuk.edu.tw

Abstract. Mining generalized association rules among items in the presence of taxonomy and with nonuniform minimum support has been recognized as an important model in the data mining community. In real applications, however, the work of discovering interesting association rules is an iterative process; the analysts have to continuously adjust the constraint of minimum support to discover real informative rules. How to reduce the response time for each remining process thus becomes a crucial issue. In this paper, we examine the problem of maintaining the discovered multi-supported generalized association rules when the multiple minimum support constraint is refined and propose a novel algorithm called RGA_MSR to accomplish the work. By keeping and utilizing the set of frequent itemsets and negative border, and adopting vertical intersection counting strategy, the proposed RGA_MSR algorithm can significantly reduce the computation time spent on rediscovery of frequent itemsets and has very good performance.

1 Introduction

Mining association rules from a large database of business data, such as transaction records, has been a popular topic within the area of data mining [1, 2, 4, 8, 10, 13, 15]. This problem is originally motivated by application known as market basket analysis to find correlations among items purchased by customers.

An association rule is an expression of the form $X \Rightarrow Y$, where X and Y are sets of items. Such a rule reveals that transactions in the database containing items in X tend to contain items in Y, and the probability, measured as the fraction of transactions containing X also containing Y, is called the *confidence* of the rule. The *support* of the rule is the fraction of the transactions that contain all items in both X and Y. For an association rule to be valid, the rule should satisfy a user-specified minimum support, called *ms*, and minimum confidence, called *mc*, respectively.

In many applications, there are taxonomies (hierarchies), explicitly or implicitly, over the items. Therefore, it may be more useful to find associations at different levels of the taxonomy rather than only at the primitive concept level [5, 11]. For example, consider Fig. 1.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1338–1344, 2005.

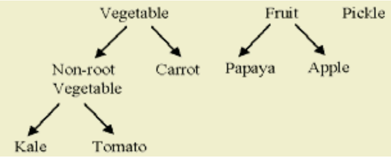


Fig. 1. An example of taxonomy *T*

It is likely to happen that the association rule

$$\text{Carrot} \Rightarrow \text{Apple (Support = 30\%, Confidence = 60\%)},$$

does not hold when the minimum support is set to 40%, but the following association rule may be valid

$$\text{Vegetable} \Rightarrow \text{Fruit}$$

In our previous work, we have investigated the problem of, and proposed two efficient algorithms, MMS_Cumulate and MMS_Stratify, for mining generalized association rules across different levels of taxonomy with multiple minimum supports [14]. In real applications, however, the work of discovering interesting association rules is an iterative process; the analysts have to continuously adjust the constraint of minimum support and/or minimum confidence to discover real informative rules. How to reduce the response time for each remining process thus becomes a crucial issue.

In this paper, we consider the updating approach and propose an algorithm, called RGA_MSR (Remining Generalized Association rules under multiple Minimum Supports Refinement). Our algorithm, by utilizing the discovered frequent itemsets and infrequent candidate itemsets in the previous mining process, can significantly reduce the number of candidate itemsets as well as database rescanning. Empirical evaluation showed that our algorithm is much more efficient than that applying our previously proposed MMS_Cumulate and MMS_Stratify algorithms afresh to accomplish the remining task.

The remaining of this paper is organized as follows. A description of related work is given in Section 2. In Section 3, we explain how to remine generalized association rules under multiple minimum support refinement and describe the proposed RGA_MSR algorithm. Evaluation of the RGA_MSR algorithm is described in Section 4. Finally, our conclusion is stated in Section 5.

2 Related Work

The problem of mining association rules in the presence of taxonomy information is first addressed in [6] and [11], independently. In [11], the authors aim at finding associations among items at any level of the taxonomy under the *ms* and *mc* constraints. In [6], they are primarily devoted to mining associations level-by-level in a fixed hierarchy, in which the uniform minimum support constraint is generalized to a form of level-wise assignment.

Another form of association rule model with multiple minimum supports is proposed in [7]. Their method allows users to specify different minimum supports to different item and can find rules involving both frequent and rare items. However, their model considers no taxonomy at all, and hence fails to find generalized association rules.

3 The Proposed Method

3.1 Algorithm Basics

Let a k-itemset denote an itemset with k items. The basic process of our generalized association rules remining algorithm under multiple minimum support update follows the level-wise approach used by most Apriori-like algorithms. However, the well-known Apriori pruning technique based on the concept of *downward closure* does not work for multiple support specification. To solve this problem, we have adopted the *sorted closure property* [7] in our previous work for mining generalized association rules with multiple minimum supports. Hereafter, to distinguish from the traditional itemset, a sorted k-itemset denoted as $\langle a_1, a_2, \ldots, a_k\rangle$ is used.

Lemma 1 If a sorted k-itemset $\langle a_1, a_2, \ldots, a_k\rangle$, for $k \geq 2$ and $ms(a_1) \leq ms(a_2) \leq \ldots \leq ms(a_k)$, is frequent, then all of its sorted subsets with k–1 items are frequent, except the subset $\langle a_2, a_3, \ldots, a_k\rangle$.

Lemma 2 For $k = 2$, the procedure apriori-gen(L_1) fails to generate all candidate 2-itemsets in C_2.

Lemma 3 For $k \geq 3$, any k-itemset $A = \langle a_1, a_2, \ldots, a_k\rangle$ generated by procedure apriori-gen(L_{k-1}) can be pruned if there exists one (k–1) subset of A, say $\langle a_{i_1}, a_{i_2}, \ldots, a_{i_{k-1}}\rangle$, such that $\langle a_{i_1}, a_{i_2}, \ldots, a_{i_{k-1}}\rangle \notin L_{k-1}$ and $a_{i_1} = a_1$ or $ms(a_{i_1}) = ms(a_{i_2})$.

For more details, please refer to [14].

3.2 Algorithm RGA_MSR

Our RGA_MSR algorithm proceeds as follows. First, we load all 1-itemsets which are frequent or are infrequent but located at the negative border (NB) with respect to the old multiple minimum supports (ms_{old}) setting. Second, we create the frontier set F and use it to generate the set of candidate 2-itemsets C_2. Next, we create the set of new frequent 1-itemsets L_1^{new} according to the new multiple minimum supports (ms_{new}) setting. We then generate the set of frequent 2-itemsets L_2^{new} by scanning some part of C_2 in DB. Finally, for $k \geq 3$, we repeatedly generate the set of candidate k-itemsets C_k from L_{k-1}^{new} and create the set of frequent k-itemsets L_k^{new} until no frequent itemsets are generated.

Lemma 4 An itemset A is frequent with respect to $ms_{new}(A)$ if it is frequent with respect to $ms_{old}(A)$ and $ms_{new}(A) \leq ms_{old}(A)$, where $ms(A) = \min_{a_i \in A} ms(a_i)$.

Lemma 5 An itemset A is infrequent with respect to $ms_{new}(A)$ if it is infrequent with respect to $ms_{old}(A)$ and $ms_{new}(A) > ms_{old}(A)$.

In addition, an itemset A is uncertain of frequency with respect to $ms_{new}(A)$ if it is frequent with respect to $ms_{old}(A)$ and $ms_{new}(A) > ms_{old}(A)$, or if it is infrequent with respect to $ms_{old}(A)$ and $ms_{new}(A) < ms_{old}(A)$.

For itemsets satisfying Lemma 4, there is no need to rescan the database DB to determine whether they are frequent. For those satisfying Lemma 5, we already know itemsets are infrequent; therefore, they are just discarded. In addition, three cases are considered: one is that the itemsets which are frequent with respect to ms_{old} and $ms_{new} > ms_{old}$ are not required to rescan the database DB, but they are required to be calculated to determine whether they are frequent, another case is that the itemsets which are infrequent with respect to ms_{old} and $ms_{new} < ms_{old}$ but found in NB are not required to rescan the database DB; however, itemsets can be determined whether they are frequent by simply comparing with ms_{new}, and the other case is that the itemsets which are infrequent with respect to ms_{old} and $ms_{new} < ms_{old}$ but not found in NB need to rescan the database DB.

The main steps of RGA_MSR Algorithm are presented as follows:

Inputs: (1) DB: the database; (2) the old multiple minimum support (ms_{old}) setting; (3) the new multiple minimum support (ms_{new}) setting; (4) T: the item taxonomy; (5) $L^{old} = \bigcup_k L_k^{old}$: the set of old frequent itemsets; (6) $NB = \bigcup_k NB_k$: the set of old infrequent candidate itemsets.

Output: $L^{new} = \bigcup_k L_k^{new}$: the set of new frequent itemsets with respect to ms_{new}.

Steps:

1. Load L_1^{old} and NB_1.
2. Divide the set of 1-itemsets C_1 into two parts: one X_1 consists of items in L_1^{old} , and the other Y_1 contains those in NB_1.
3. Sort C_1 in increasing order of their mss, and create frontier set F and L_1^{new} .
4. Generate the set of candidate 2-itemsets C_2 from F.
5. Add generalized items in T into DB as ED, and convert ED into tidlist and bitmap.
6. Delete any candidate in C_2 that consists of an item and its ancestor.
7. Load L_2^{old} and NB_2.
8. Divide C_2 into three parts: the itemsets in L_2^{old} , those in NB_2, and those not in NB_2. For the itemsets in L_2^{old} , further divide them into two parts: X_{2a} for $ms_{new} \leq ms_{old}$ and X_{2b} for $ms_{new} > ms_{old}$. For those in NB_2, further divide them into two parts: Y_{2a} for $ms_{new} < ms_{old}$ and Y_{2b} for $ms_{new} \geq ms_{old}$. For those not in NB_2, further divide them into two parts: Z_{2a} for $ms_{new} < ms_{old}$ and Z_{2b} for $ms_{new} \geq ms_{old}$.
9. Count the supports of itemsets in Z_{2a} over tidlist and bitmap.
10. Create L_2^{new} by combining X_{2a} and those itemsets which are frequent in X_{2b}, Y_{2a} and Z_{2a}.
11. Generate candidates C_3 from L_2^{new} .
12. Repeat Steps 6-11 for new candidates C_k until no frequent k-itemsets L_k^{new} are created.

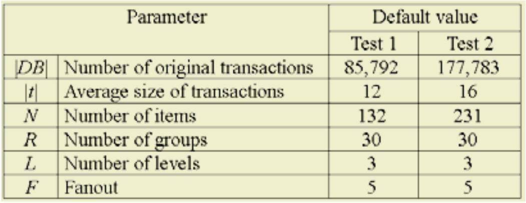
An example illustrating the RGA_MSR algorithm is provided in Appendix, where in Fig. 4, item “A” stands for “Vegetable”, “B” for “Non-root Vegetable”, “C” for “Kale”, “D” for “Carrot”, “E” for “Tomato”, “F” for “Fruit”, “G” for “Papaya”, “H” for “Apple”, and “I” for “Pickle”.

4 Experiments

In this section, we describe the experiments conducted on evaluating the performance of algorithm RGA_MSR using two synthetic datasets generated by IBM data generator [2]. In the implementation of RGA_MSR, we have kept and utilized the set of frequent itemsets and negative border [3, 9, 12, 13], and adopted the vertical intersection counting strategy [1, 2, 4, 8, 10, 15]. The parameter settings for the two test sets are shown in Table 1. All experiments were performed on an Intel Pentium-IV 2.80GHz with 2GB RAM, running on Windows 2000.

First, the efficiency of RGA_MSR was compared to MMS_Stratify and MMS_Cummulate under various sizes of databases. The minimum supports were specified to items randomly, ranging from 0.4% to 5.5%. Besides, we adopted the ordinary case that the minimum support of an item a is no larger than any of its ancestors $\hat{a}$, i.e., $ms(a) \leq ms(\hat{a})$. The comparison was performed under two different cases: 1) all items having $ms_{new} > ms_{old}$; and 2) all items having $ms_{new} < ms_{old}$. As the results depicted in Fig. 2 and 3 show, RGA_MSR significantly outperforms MMS_Stratify and MMS_Cumulate.

Table 1. Parameter settings for synthetic data



Parameter		Default value	
		Test 1	Test 2
$\|DB\|$	Number of original transactions	85,792	177,783
$\|t\|$	Average size of transactions	12	16
N	Number of items	132	231
R	Number of groups	30	30
L	Number of levels	3	3
F	Fanout	5	5

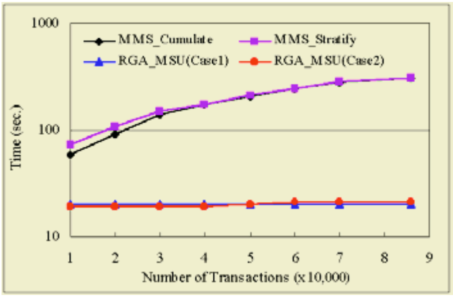


Fig. 2. Execution time for various sizes of transactions for test set 1

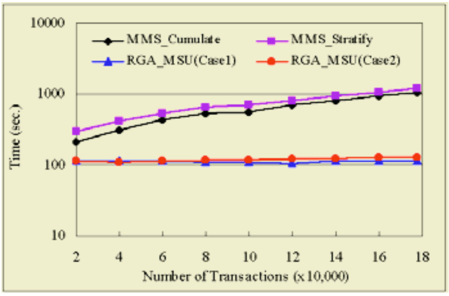


Fig. 3. Execution time for various sizes of transactions for test set 2

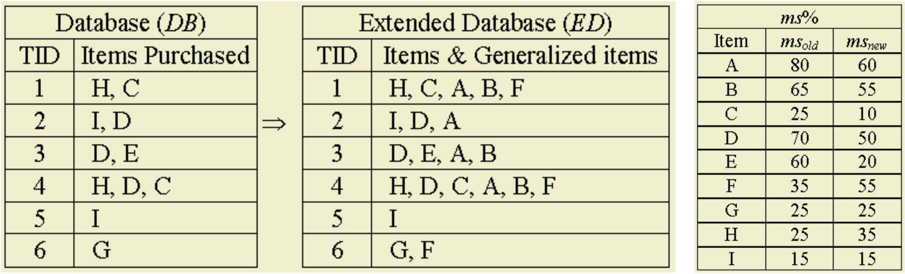
5 Conclusions

We have investigated in this paper the problem of remining association rules under multiple minimum support refinement and presented an efficient algorithm, RGA_MSR. Empirical evaluation showed that the algorithm is very efficient to complete the remining process under multiple minimum support refinement, and superior to our previous proposed algorithms MMS_Cumulate and MMS_Stratify. In the future, we will continue this study to find more efficient algorithms. We will also consider applying our method to the problem of on-line discovery and maintenance of multi-dimensional association rules from data warehouses.

References

1. Agrawal R., Imielinski T., Swami A.: Mining Association Rules between Sets of Items in Large Databases. In Proc. 1993 ACM-SIGMOD Intl. Conf. Management of Data (1993) 207-216
2. Agrawal R., Srikant R.: Fast Algorithms for Mining Association Rules. In Proc. 20th Intl. Conf. Very Large Data Bases (1994) 487-499
3. Ayan N.F., Tansel A. U., Arkun M.E.: An Efficient Algorithm to Update Large Itemsets with Early Pruning. In Proc. of the 5th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD 1999), San Diego, CA, USA (1999) 287-291
4. Brin S., Motwani R., Ullman J.D., Tsur S.: Dynamic Itemset Counting and Implication Rules for Market Basket Data. SIGMOD Record, Vol. 26 (1997) 255-264
5. Cheung D.W., Lee S.D., Kao B.: A General Incremental Technique for Maintaining Discovered Association Rules. In Proc. DASFAA'97 (1997) 185-194
6. Han J., Fu Y.: Discovery of Multiple-level Association Rules from Large Databases. In Proc. 21st Intl. Conf. Very Large Data Bases, Zurich, Switzerland (1995) 420-431
7. Liu B., Hsu W., Ma Y.: Mining Association Rules with Multiple Minimum Supports. In Proc. 5th Intl. Conf. Knowledge Discovery and Data Mining (1999) 337-341
8. Park J.S., Chen M.S., Yu P.S.: An Effective Hash-based Algorithm for Mining Association Rules. In Proc. 1995 ACM SIGMOD Intl. Conf. on Management of Data, San Jose, CA, USA (1995) 175-186
9. Sarda N.L., Srinivas N.V.: An Adaptive Algorithm for Incremental Mining of Association Rules. In Proc. 9th Intl. Workshop on Database and Expert Systems Applications, (1998) 240-245
10. Savasere A., Omiecinski E., Navathe S.: An Efficient Algorithm for Mining Association Rules in Large Databases. In Proc. 21st Intl. Conf. Very Large Data Bases (1995) 432-444
11. Srikant R., Agrawal R.: Mining Generalized Association Rules. In Proc. 21st Intl. Conf. Very Large Data Bases (1995) 407-419
12. Thomas S., Bodagala S., Alsabti K., Ranka S.: An Efficient Algorithm for the Incremental Updation of Association Rules in Large Databases. In Proc. 3rd Intl. Conf. Knowledge Discovery and Data Mining (1997)
13. Toivonen H.: Sampling Large Databases for Association Rules. In Proc. 22nd Intl. Conf. on Very Large Data Bases, Mumbai (India) (1996) 134-145
14. Tseng M.C., Lin W.Y.: Mining Generalized Association Rules with Multiple Minimum Support. In Proc. Intl. Conf. Data Warehousing and Knowledge Discovery (2001) 11-20
15. Zaki M.J.: Scalable Algorithms for Association Mining. IEEE Transactions on Knowledge and Data Engineering, Vol. 12, No. 2 (2000) 372-390

Appendix



Database (*DB*)	
TID	Items Purchased
1	H, C
2	I, D
3	D, E
4	H, D, C
5	I
6	G

⇒

Extended Database (*ED*)	
TID	Items & Generalized items
1	H, C, A, B, F
2	I, D, A
3	D, E, A, B
4	H, D, C, A, B, F
5	I
6	G, F

	ms%	
Item	ms_{old}	ms_{new}
A	80	60
B	65	55
C	25	10
D	70	50
E	60	20
F	35	55
G	25	25
H	25	35
I	15	15

Fig. 4. An example for illustration of RGA_MSR

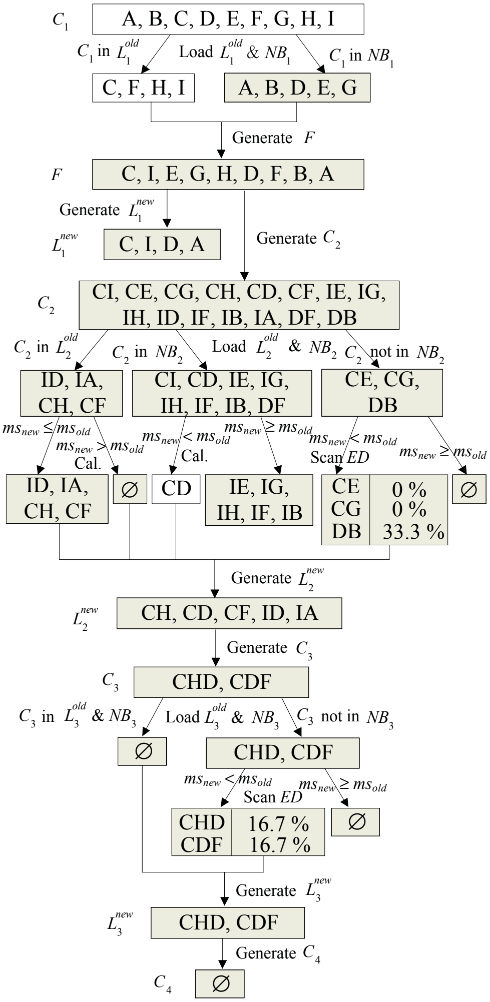


Fig. 5. Illustration of algorithm RGA_MSR

Mining Association Rules from Distorted Data for Privacy Preservation*

Peng Zhang[1,2], Yunhai Tong[1,2], Shiwei Tang[1,2], and Dongqing Yang[1,2]

[1] School of EECS, Peking University, Beijing, 100871, China
[2] National Lab on Machine Perception, Peking University, Beijing, 100871, China
{zhangpeng,yhtong,tsw,ydq}@db.pku.edu.cn

Abstract. In order to improve privacy preservation and accuracy, we present a new association rule mining scheme based on data distortion. It consists of two steps: First, the original data are distorted by a new randomization method. Then, the mining algorithm is implemented to find frequent itemsets from the distorted data, and generate association rules. With reasonable selection for the random parameters, our scheme can simultaneously provide a higher privacy preserving level to the users and retain a higher accuracy in the mining results.

1 Introduction

Association rule mining is to find interesting relationships among a large set of data items. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining association rules, but privacy issues are also brought to them simultaneously. For example, a hospital could mine the cases of patients to find the association relationships among all kinds of diseases, but it is inescapable to uncover the patients' data so as to breach their personal privacy by the ordinary mining approaches. Therefore, how to solve the privacy preserving problem during mining process has become one of the most important topics in data mining [1].

However, it is not the technology of data mining itself but the methods adopted that possibly breach the privacy. Data mining has an essential property that the patterns from large amounts of data usually depend on the aggregate and statistical data but not the individual data records. Privacy preserving data mining is to discover accurate patterns without precise access to the original data.

In order to improve the privacy preservation and accuracy for mining association rules, we present a new scheme consisting of two steps. First, a new randomization method is proposed to distort the original data. Then, the mining algorithm on the distorted data is implemented to find frequent itemsets, and generate association rules.

The remainder of this paper is organized as follows: In section 2, we introduce the related work. Section 3 presents a novel approach to mining association rules from distorted data. The analytical and experimental results are shown in section 4. Finally, in section 5, we summarize the conclusions and outline future work.

* This work is supported by the National Natural Science Foundation of China under Grant No.60403041.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1345–1351, 2005.

2 Related Word

The proposed approaches in privacy preserving data mining can be classified into the following two categories: distortion based approach and partition based approach.

The central idea of distortion based approach is that we first distort the values of the original data using data transform, data hiding, and adding noise. Then, the underlying patterns of the original data are discovered by reconstruction techniques.

MASK scheme was presented based on probabilistic distortion [2], and optimized in reconstruction process [3]. Recovering association rules from uniform randomized data by computing the partial supports was also proposed [4]. However, the distorted data are all directly transformed from the original data, and the random parameters are constrained to be not close to 0.5. Besides the above approaches using data transform, a method for selectively removing individual values from a database was presented to prevent the discovery of a set of sensitive rules [5]. Then, [6] also proposed to hide sensitive rules by reducing their supports. Although a part of sensitive information is protected well, the supplied data for mining are all original so that the privacy preservation for the whole data set is not satisfactory. Furthermore, the sensitive rules to be hidden must be known before mining, which is usually impossible.

The partition based approach is to use the Secure Multiparty Computation (SMC) techniques. Each data mining participant just has a part of the original data, and then the patterns are developed by a set of secure protocols for distributed computation.

Privacy preserving association rule mining for horizontally partitioned data [7] that one party just has a set of records and vertically partitioned data [8] that one party just has some of the attributes were respectively presented. However, these approaches are meaningful only in the context of distributed database. All the data suppliers must participate in mining process. One party fault may lead to error results, even mining failure, and the performance is not desirable when the party count becomes large.

3 Mining Association Rules from Distorted Data

This section presents a novel approach to mining association rules from distorted data.

3.1 Framework

Let $I=\{i_1, i_2, ..., i_m\}$ be a set of items. D is a set of transactions. Each transaction $T\subseteq I$, is represented by a vector, $T=(t_1, t_2, ..., t_m)$, where $t_j\in\{0, 1\}$, j=1, 2, ..., m, respectively describes whether T contains the item i_j. If T contains i_j, t_j=1; otherwise, t_j=0. An association rule is an implication of the form $X\Rightarrow Y$, where $X\subset I$, $Y\subset I$, and $X\cap Y=\Phi$. The rule $X\Rightarrow Y$ holds with support, the percentage of transactions that contain both X and Y. The rule $X\Rightarrow Y$ has confidence that is the percentage of transactions containing X also contain Y. Association rule mining is a two step process:

1. Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a pre-determined minimum support count.
2. Generate strong association rules from the frequent itemsets.

In this paper, we present a privacy preserving association rule mining scheme including two steps. The framework is shown in Fig. 1. In the first step, we distort the original data by a new randomization method. In the second step, we find the frequent itemsets from the distorted date, and then generate the strong association rules.

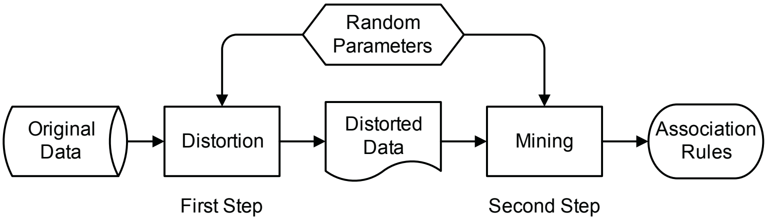


Fig. 1. Framework

3.2 Data Distortion Method

We make use of Randomized Response Techniques (RRT) to distort the original data. The essential idea of RRT is that the original data are first distorted by the data supplier according to the random parameters, and then provided to the data user. Although, the detailed original data are distorted, the statistical and aggregate information can still be precisely estimated while there are large amounts of data. Association rule mining depends on the aggregate information of a data set, but not the individual data records, so RRT can be well used.

However, RRT used in the existing privacy preserving data mining approaches are all based on Warner's model. Not only are all the distorted data directly transformed from the original data so as to degrade the privacy preserving level, but also the random parameters are constrained to be not close to 0.5.

The strategy of data hiding can overcome the above disadvantage, but bring another one that the supplied data are all exactly the original data. How about integrating the above two strategies for data distortion to improve the accuracy of mining with better privacy preservation?

Just as this idea, we combine the two strategies of data transform and data hiding to propose a new randomization method. During the distortion process, the original data are simultaneously transformed and hidden. There is also no constraint to select the random parameters. The detailed description of our method is as follows:

Given a set of random parameters, $0\leq p_1, p_2, p_3 \leq 1$, such that $p_1+p_2+p_3=1$. For each $t\in\{0,1\}$, let $r_1=t$, $r_2=1$, $r_3=0$, the random function $r(x)$ returns the value r_i with probability p_i (i=1, 2, 3). A transaction in D can be considered to be a vector $T=(t_1, t_2, ..., t_m)$, such that $t_i\in\{0,1\}$. We generate the distorted vector $T'=(t'_1, t'_2, ..., t'_m)$ by computing $T'=R(T)$, where $t'_j=r(t_j)$. That is, t'_j takes a value t_j with probability p_1, 1 with probability p_2, and 0 with probability p_3.

In this way, each transaction T in the original transaction set D can become a distorted transaction T' by the random function $T'=R(T)$. Since the form of T' is still similar to an original transaction, it can be added into the distorted transaction set D' as a disguised transaction.

In principle, it is possible to use different sets of random parameters for distorting different items. For simplicity, we will assume here that a single set of random parameters, p_1, p_2, p_3, for all the items.

3.3 Data Mining Approach

The original transaction set D has become a distorted transaction set D'. Then, the process to find frequent itemsets and generate association rules is described in this section. We first present the method for estimating the supports of 1-itemsets, and then show how to estimate the supports of k-itemsets.

Given an item i, let π be its support in D, and λ be its support in D'. Suppose a transaction T becomes a distorted transaction T'. Then, the values of t_i, t'_i, and their mapping probabilities are shown in Table 1.

Table 1. Probabilities of data mapping by data distortion

No.	t_i	t'_i	Probability
1	0	0	p_1+p_3
2	0	1	p_2
3	1	0	p_3
4	1	1	p_1+p_2

From Table 1, we can obtain $\lambda = \pi(p_1+p_2)+(1-\pi)p_2 = \pi p_1+p_2$. Then, $\pi = (\lambda-p_2)/p_1$. We first compute λ, and then π is to be estimated by the above equation.

Let $A=\{i_1, i_2, ..., i_k\}$ be a k-itemset. There are k+1 possible values for $|T\cap A|$, where T is a transaction in D. Each proportion of transactions in D containing the same count of items in A is respectively represented as C_0, C_1, ..., C_k. Analogously, for a distorted transaction T' in D', there are also k+1 possible values for $|T'\cap A|$. C'_0, C'_1, ..., C'_k respectively represent the proportions of transactions in D' containing the same count of items in A. When we select the same random parameters for all the items, the probabilities, m_{ij}, that an original transaction in D having j items of A becomes a distorted transaction in D' having i items of A are uniform for each pair of i, j $(0 \le i, j \le k)$, and can be computed by the equation (1).

$$m_{ij} = \sum_{t=\max(0,i+j-k)}^{\min(i,j)} C_j^t \cdot (p_1+p_2)^t \cdot p_3^{j-t} \cdot C_{k-j}^{i-t} \cdot p_2^{i-t} \cdot (p_1+p_3)^{k-i-j+t} \tag{1}$$

Thus, $C'=MC$, where $C' = \begin{bmatrix} C'_0 \\ C'_1 \\ \vdots \\ C'_k \end{bmatrix}$, $C = \begin{bmatrix} C_0 \\ C_1 \\ \vdots \\ C_k \end{bmatrix}$, and $M=[m_{ij}]$ is a $(k+1)'(k+1)$ matrix.

When M is invertible, let $M^{-1}=[a_{ij}]$, $C=M^{-1}C'$. C_k is just the support of the k-itemset that we need estimate. We first count each C'_j from the distorted transaction set D', and compute a_{kj} from M. Then, the support of k-itemset A in the original transaction set D can be estimated by the equation (2).

$$C_k = a_{k0}\cdot C'_0 + a_{k1}\cdot C'_1 + ... + a_{kk}\cdot C'_k \tag{2}$$

Additionally, note that $C'_0+ C'_1+...+ C'_k=|D'|=N$. Therefore, one of all the C'_j can be computed without counting. Usually, the value of C'_0 is much more that any other values. So, the value of C'_0 can be computed by $C'_0=N-(C'_1+...+ C'_k)$.

4 Analysis and Experiments

In this section, our scheme is amply analyzed and compared with MASK scheme [2] in terms of privacy preservation and mining accuracy. Then, experiments are used to further verify the effectiveness of our scheme.

4.1 Analysis

Just as its name implies, the privacy preserving level is the most important measure to evaluate the schemes. In order to quantitatively compare the schemes better, we define a privacy measure, *Breach*, called privacy breach coefficient.

$$Breach = P_{\text{true}} \cdot P_{\text{identifying true}} + P_{\text{false}} \cdot P_{\text{identifying false}} \cdot P_{\text{reconstruction}} \tag{3}$$

Here, P_{true} and P_{false} are respectively the proportions of true and false data in the distorted data set, such that $P_{\text{true}}+P_{\text{false}}=1$. $P_{\text{identifying true}}$ is the probability for identifying true data from the distorted data set, while $P_{\text{identifying false}}$ is the probability for identifying false data from the distorted data set. $P_{\text{reconstruction}}$ is the probability with which the false data can be reconstructed to the relative original.

Now, we compare the privacy preserving levels of MASK scheme and our scheme by respectively computing their privacy breach coefficients. Suppose the proportions of true original data in the distorted data sets are same for the two schemes. Then,

1. MASK scheme, the random parameter is p. $Breach_1=p\cdot p+(1-p)\cdot(1-p)\cdot 1=p^2+(1-p)^2$.
2. Our scheme, $p_1=p$, $p_2=p_3=\frac{1-p_1}{2}$, $Breach_2= p_1 \cdot \frac{p_1}{p_1+p_2} = \frac{2p^2}{p+1}$.

The relation between $Breach_1$ and $Breach_2$ depends on the value of the random parameter p. $\Delta_1=Breach_1-Breach_2=\frac{2p^3-2p^2-p+1}{p+1}=\frac{(\sqrt{2}p+1)(\sqrt{2}p-1)(p-1)}{p+1}$.

Thus, when $0<p<\frac{1}{\sqrt{2}}$, the privacy breach coefficients meet $Breach_1 > Breach_2$.

Privacy and accuracy seem to be contradictive. The privacy preservation must lead to the descent of accuracy. In order to increase the accuracy, the privacy preserving level must be declined as the cost. The above analysis has made it out that our scheme can obtain a higher privacy preserving level than MASK scheme. Now, we further explain by variance analysis that our scheme can simultaneously make progress in terms of accuracy, only if the random parameters are appropriately selected.

We also assume the proportions of true original data in the distorted data set are same for the two schemes. Then,

1. MASK scheme, $\hat{\pi}_1=\frac{\lambda_1-(1-p)}{2p-1},(p\neq\frac{1}{2})$, variance $Var(\hat{\pi}_1)=\frac{p(1-p)}{n(2p-1)^2},(p\neq\frac{1}{2})$.

2. Our scheme, $\hat{\pi}_2=\frac{\lambda_2-p_2}{p_1}$, variance $Var(\hat{\pi}_2)=\frac{p_1(1-p_1)\pi}{np_1^2}+\frac{p_2(1-p_2)-2\pi p_1 p_2}{np_1^2}$.

   For $p_1=p$, $p_2=p_3=\frac{1-p_1}{2}$, then $Var(\hat{\pi}_2)=\frac{p_2(1-p_2)}{np_1^2}=\frac{(1-p)(1+p)}{4np^2}$.

   Both $\hat{\pi}_1$ and $\hat{\pi}_2$ are enormous likelihood unbiased estimator for π, but

   $$\Delta_2=Var(\hat{\pi}_1)-Var(\hat{\pi}_2)=\frac{1-p}{n}\left[\frac{p}{(2p-1)^2}-\frac{1+p}{4p^2}\right]=\frac{(1-p)(3p-1)}{4np^2(2p-1)^2},(p\neq\frac{1}{2}).$$

   Therefore, when $\frac{1}{3}<p<1$, the variances meet $Var(\hat{\pi}_1)>Var(\hat{\pi}_2)$.

Combined with the above analysis, we come to the conclusion: when $\frac{1}{3}<p<\frac{1}{\sqrt{2}}$, our scheme can simultaneously provide higher privacy preserving level and higher accuracy than the existing MASK scheme. By the trade-off between privacy preservation and accuracy, this proportion scope of true original data in the distorted data set is satisfactory in practice. Furthermore, the constraint to select the random parameter, i.e. $p\neq 0.5$, in MASK scheme is also avoided in our scheme.

4.2 Experiments

Our experiments were carried out on a synthetic data set that was generated by the IBM Almaden generator with the parameters *T*10.*I*4.*D*100k.*N*100. MASK scheme is implemented as a contrast. The relationships of privacy preservation, mining accuracy, and the random parameters are also explained.

Let F be the set of frequent itemsets in D. f is a frequent itemset in F, and its actual support is s_f. The support of f estimated by a privacy preserving mining scheme from the distorted transaction set D' is s'_f. Then, the support error of f is defined as $SE_f=\frac{|s_f-s'_f|}{s_f}$. The total support error of the scheme is defined as $SE=\frac{1}{|F|}\sum_f SE_f$.

We selected different random parameters $p_1=p=0.1$, 0.2, 0.3, 0.35, 0.4, 0.45, 0.49, 0.51, 0.55, 0.6, 0.65, 0.7, 0.8, 0.9; $p_2=p_3=(1\text{-}p_1)/2$. Total support errors of our scheme and MASK scheme were respectively computed for different minimum supports.

Fig. 2 shows the average support errors of MASK scheme and our scheme for each different p values. When p is very small, MASK scheme is more accurate than our scheme; while p is over 0.3, the accuracy of our scheme is increasing beyond MASK scheme. Especially, when p is around 0.5, it can exceed MASK scheme much more.

When p is moving from 0 to 1, the accuracy of our scheme is continually enhancing, but the privacy preserving level is coming down. However, for MASK scheme, when p is near to 0 or 1, the result is quite accurate but the privacy preserving level is fairly low. As p is away from 0 or 1, and approaches to 0.5, the privacy preserving level is increasing, but the accuracy is remarkably descending.

According to the analytical and experimental results, we propose to select p from the interval [0.35, 0.6] to use our scheme for privacy preserving association rule mining by the trade-off between privacy preservation and accuracy.

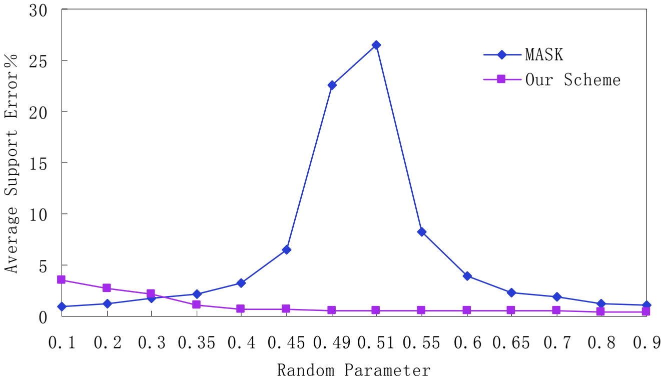


Fig. 2. Average Support Errors

5 Conclusions and Future Work

In this paper, we have presented a privacy preserving association rule mining scheme consisting of two steps: data distortion and mining from the distorted data. Our scheme has been analyzed in terms of privacy preservation and mining accuracy. With reasonable selection for the random parameters, it can provide a higher privacy preserving level to the users and retain a higher accuracy in the mining results.

In the future work, we will apply our scheme to solve other data mining problems. We will also extend our randomization and reconstruction method to support privacy preserving mining categorical data.

References

1. V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis. State-of-the-art in Privacy Preserving Data Mining. In SIGMOD Record, 33(1), 2004.
2. S. J. Rizvi and J. R. Haritsa. Maintaining Data Privacy in Association Rule Mining. In Proceedings of the 28th International Conference on Very Large Data Bases, 2002.
3. S. Agrawal, V. Krishnan and J. R. Haritsa. On Addressing Efficiency Concerns in Privacy-Preserving Mining. In Proceedings of the 9th International Conference on DASFAA, 2004.
4. A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy Preserving Mining of Association Rules. Proc. of the 8th ACM SIGKDD International Conference on KDD, 2002.
5. Y. Saygin, V. S. Verykios and C. Clifton. Using Unknowns to Prevent Discovery of Association Rules. ACM SIGMOD Record, 30(4), 2001.
6. S. R. M. Oliveira and O. R. Za1ane. Privacy Preserving Frequent Itemset Mining. In Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining, 2002.
7. M. Kantarcioglu and C. Clifton. Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data. IEEE Transactions on Knowledge and Data Engineering 16(9), 2003.
8. J. Vaidya and C. Clifton. Privacy Preserving Association Rule Mining in Vertically Partitioned Data. Proc. of the 8th ACM SIGKDD International Conference on KDD, 2002.

Mining Linguistic Mobility Patterns for Wireless Networks

Tzung-Pei Hong[1], Cheng-Ming Huang[2], and Shi-Jinn Horng[2]

[1] Department of Electrical Engineering, National University of Kaohsiung
National Taiwan University of Science and Technology
tphong@nuk.edu.tw
[2] Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
apcmhl@ms8.hinet.net, horng@mouse.ee.ntust.edu.tw

Abstract. Wireless networks and mobile applications have grown very rapidly and have made a significant impact on computer systems. Especially, the usage of mobile phones and PDA is increased very rapidly. Added functions and values with these devices are thus greatly developed. If some regularity can be known from the user mobility behavior, then these functions and values can be further expanded and used intelligently. This paper thus attempts to mine appropriate linguistic mobility patterns for being used by mobile-system managers in future strategy planning. The location areas in which mobile users visit and their duration times can be found from the log data stored in the home-location-register module. Since the duration times are numeric, fuzzy concepts are used to process them and to form linguistic mobility patterns.

1 Introduction

Wireless networks and mobile applications have recently become very popular and have made a significant impact on our daily life [3]. Especially, the usage of mobile phones and PDA is increased very rapidly. Added functions and values with these devices are thus greatly developed. In a wireless system of mobile phones, a module called home location register (HLR) keeps related user mobility data. When a mobile user moves from a location area to its next one, his/her corresponding data in HLR are updated. HLR can thus help the system successfully and efficiently find users' current locations and distribute desired messages to them. The log data stored in HLR can also be used for analyzing users' mobility behavior. The regular patterns found from the mobility data may provide some appropriate suggestions to mobile-system managers. The added functions and values with the system may also be further expanded.

Data-mining has become a process of considerable interest in recent years as the amounts of data in many databases have grown tremendously large. It has been developed to turn data into useful task-oriented knowledge. It is used to identify effective, coherent, potentially useful, and previously unknown patterns in large databases [5]. Some data-mining approaches for mobile systems were proposed in the past [12, 13]. Most uses binary-value data in the mining process. In these years, fuzzy set theory is being used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning [16, 17]. It can provide linguistic representation and

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1352–1358, 2005.

is thus natural and understandable to human beings. Several fuzzy learning algorithms for inducing rules from given sets of data have been designed and used to good effect with specific domains. Using fuzzy sets in data mining has also been developed in recent years [8, 9].

In this paper, we emphasize on the automatic discovery of linguistic user mobility patterns in wireless networks. The preprocessing task for preparing appropriate data from HLR is first done. The duration time of each location area for a mobile user to stay is calculated from the time interval between the current record and its next one and is used to analyze the mobility behavior of the mobile system. The fuzzy concepts are then used to process them and to form linguistic terms. Our previous fuzzy mining algorithm [9] is modified to find mobility patterns from these terms and log data.

2 Review of Wireless Networks for Mobile Users

The architecture of a typical wireless communication system for mobile users is shown in Figure 1 [2, 14].

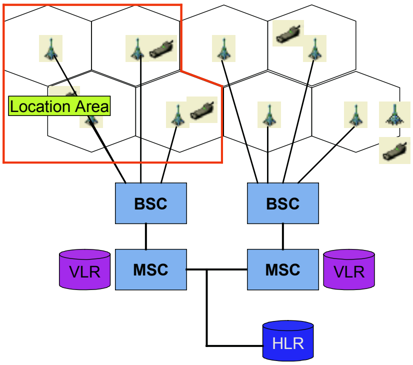


Fig. 1. A typical wireless communication system

All the communication space of a wireless system is divided into many units called cells. Each cell has a base station, in charge of broadcasting the contents to mobile users located within the cell. Several neighboring cells are grouped into a location area, which are managed by a base-station controller (BSC) and a mobile switch service center (MSC). BSC is used to control all the base stations in a location area for performing their jobs; MSC maintains a visitor location register (VLR) which stores the location area identifier and some related information to the current visitors in the location area. All the visitors with their relevant VLR data are sent to the home location register (HLR). When calls come, HLR determines users' current location areas according to the identification number of mobile users called, and invokes corresponding BSCs to broadcast messages from base stations to them. The action is called paging. Besides, when mobile users move from a location area to another one, they must register and update both the VLR data in MSC and the HLR data. This action is called location update. Many researches were thus proposed in the past for making a good trade-off between the costs of these two actions [6, 7, 10, 11, 14].

The location areas in which mobile users visit and their duration times can be easily found from the mobility data stored in VLR and HLR. This paper thus attempts to mine appropriate linguistic mobility patterns from these data for being used by mobile-system managers in future strategy planning.

3 Review of Related Mining Approaches

Agrawal and Srikant proposed a mining algorithm to discover sequential patterns from a set of transactions [1]. Five phases are included in their approach. In the first phase, the transactions are sorted first by customer ID as the major key and then by transaction time as the minor key. This phase thus converts the original transactions into customer sequences. In the second phase, the set of all large itemsets are found from the customer sequences by comparing their counts with a predefined support parameter . This phase is similar to the process of mining association rules. Note that when an itemset occurs more than one time in a customer sequence, it is counted once for this customer sequence. In the third phase, each large itemset is mapped to a contiguous integer and the original customer sequences are transformed into the mapped integer sequences. In the fourth phase, the set of transformed integer sequences are used to find large sequences among them. In the fifth phase, the maximally large sequences are then derived and output to users.

As to fuzzy mining, Hong *et al.* proposed a fuzzy mining algorithm to mine fuzzy rules from quantitative data [8]. They transformed each quantitative item into a fuzzy set and used fuzzy operations to find fuzzy rules. They also proposed a fuzzy mining approach to find web browsing patterns [9]. Cai *et al.* proposed weighted mining to reflect different importance to different items [4]. Each item was attached a numerical weight given by users. Weighted supports and weighted confidences were then defined to determine interesting association rules. Yue *et al.* then extended their concepts to fuzzy item vectors [15].

4 Mining Linguistic Mobility Patterns

In this paper, the log data stored in HLR in a wireless network are used to analyze the mobility patterns on that system. The four attributes mobile-ID, date, time, and location-area are used in the mining process. The log data to be analyzed are sorted first in the order of mobile-ID and then in the order of date and time. The duration time of each location area for a mobile user to stay can thus be calculated from the time interval between the current record and its next one. Since the duration times are numeric, fuzzy concepts are used to process them and to form linguistic terms. Our previous fuzzy mining algorithm proposed for finding web browsing patterns [9] is then modified and used to find mobility patterns from the data. The used fuzzy mining algorithm first transforms each quantitative value into a fuzzy set of linguistic terms using membership functions. The algorithm then calculates the scalar cardinality of each linguistic term on all mobile users. The mining process based on fuzzy counts is then performed to find fuzzy mobility patterns. The details of the proposed mobility-pattern mining algorithm are described as follows.

The mobility-pattern mining algorithm:

INPUT: A set of log data in HLR, a set of membership functions, a predefined minimum support value α.

OUTPUT: A set of linguistic mobility patterns

STEP 1: Extract the log data with only the four attributes *mobile-ID*, *date*, *time*, and *location-area*; denote the resulting log data as *D*.

STEP 2: Transform the *mobile-ID*s into contiguous integers (called encoded mobile ID) for convenience, according to their first turn-on time. Note that the same *mobile-ID* with two turn-off connections is given two integers.

STEP 3: Sort the resulting log data first by encoded mobile ID and then by date and time.

STEP 4: Calculate the duration times of the location areas in which each encoded mobile ID stay according to the time intervals between each current location area and its next one.

STEP 5: Form a mobility sequence D_i for each mobile user c_i by sequentially listing his/her n_i pairs of (location area, duration time), where n_i is the number of location areas visited by c_i.

STEP 6: Use our previous fuzzy mining algorithm in [9] to find the final linguistic mobility patterns.

5 An Example

In this section, a simple example is given to show how the proposed algorithm can be used to generate linguistic mobility patterns for mobile users' moving behavior according to the log data in HLR. Assume the location areas in a wireless network are shown in Figure 2. There are five location areas *A*, *B*, *C*, *D* and *E* in this example.

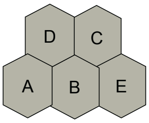


Fig. 2. The location areas used in this example

Also assume the fuzzy membership functions for the duration time of a mobile user staying in a location area are shown in Figure 3.

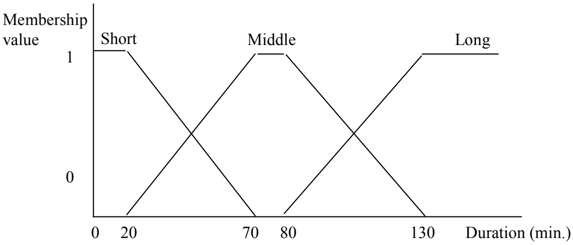


Fig. 3. The membership functions used in this example

In Figure 3, the duration time for staying in a location area is divided into three fuzzy regions: *Short, Middle* and *Long*. Thus, three fuzzy membership values are produced for each duration time according to the predefined membership functions.

Each transaction in the log data of HLR includes the four fields: *mobile-id, date, time,* and *location-area*, among others. Only the data in the above four fields are extracted for mining. Assume the extracted log data in HLR are shown in Table 1.

Table 1. The extracted log data in HLR

Mobile-id	Date	Time	Location-area
0918111111	2005-04-01	08:56	E
0918111111	2005-04-01	09:38	B
0918222222	2005-04-01	09:52	D
0918111111	2005-04-01	10:08	D
0918333333	2005-04-01	10:30	A
0918222222	2005-04-01	10:54	B
0918222222	2005-04-01	11:25	D
0918111111	2005-04-01	11:46	C
0918333333	2005-04-01	12:02	D
0918444444	2005-04-01	12:46	B
0918444444	2005-04-01	13:06	C
0918222222	2005-04-01	13:07	Turn-off
0918111111	2005-04-01	13:17	Turn-off
0918333333	2005-04-01	13:31	Turn-off
0918444444	2005-04-01	14:47	E
0918555555	2005-04-01	15:46	D
0918444444	2005-04-01	16:45	B
0918555555	2005-04-01	16:53	B
0918444444	2005-04-01	16:56	C
0918555555	2005-04-01	17:19	C
0918444444	2005-04-01	17:38	Turn-off
0918666666	2005-04-01	17:53	D
0918555555	2005-04-01	18:33	Turn-off
0918666666	2005-04-01	19:13	C
0918666666	2005-04-01	20:14	E
0918666666	2005-04-01	22:16	B
0918666666	2005-04-01	22:33	Turn-off

In this example, six mobile users logged in HLR. Only one location area is contained in each transaction. The values in the field mobile-id are then transformed into contiguous integers according to the order of each mobile user's first visiting time. The extracted log data in Table 1 are sorted first by encoded client ID and then by date and time. The duration times of the location areas visited by each mobile user are then calculated. Take the first location area by the first mobile user as an example. He enters the location-area E at 08:56 on April 1st, 2005 and the next location-area B at 09:38. The duration time of E for the first mobile user is then 42 minutes (2005/04/01, 09:38 -

2005/04/01, 08:56). After that, all the location areas visited by each mobile user are listed as a mobility sequence. The resulting mobility sequences are shown in Table 2.

Table 2. The mobility sequences formed from Table 1.

Encoded Mobile-id	Mobility Sequence
1	(E, 42) (B, 30) (D, 98) (C, 91)
2	(D, 62) (B, 31) (D, 102)
3	(A, 92) (D, 89)
4	(B, 20) (C, 101) (E, 118) (B, 11)(C, 42)
5	(D, 64) (B, 29) (C, 74)
6	(D, 80)(C, 61) (E, 122)(B, 17)

Use the fuzzy mining algorithm proposed in [9], the large 2-sequences found in this example are *(B.Short, C.Middle)*, *(D.Middle, B.Short)* and *(D.Middle, C.Middle)*. There are no large 3-sequences in this example. These three linguistic mobility patterns are then output as meta-knowledge concerning the given log data in HLR.

6 Conclusions

In this paper, we have attempted to discover linguistic mobility patterns from the extracted log data in HLR. The duration time of each location area visited by a mobile user is calculated from the time interval between the location area and its next one. Since the duration times are numeric, fuzzy concepts are used to process them and to form linguistic terms. A fuzzy mining process has then been performed to find fuzzy mobility patterns. The mined patterns are expressed in linguistic terms, which are more natural and understandable for human beings. The linguistic mobility patterns may also be used by mobile-system managers in future strategy planning, such as for location size re-planning. For example, if a linguistic mobility pattern (A.long -> B.short) is mined, the location area B may be considered to be merged into the location A. Also, message distribution strategy may also be decided based on these patterns. In the future, we will attempt to design other mining models for solving various mobile problems.

Acknowledgment

This research was supported by the National Science Council of the Republic of China under contract NSC93- 2213-E-390-001.

References

1. R. Agrawal, R. Srikant: “Mining Sequential Patterns”, *The Eleventh International Conference on Data Engineering,* 1995, pp. 3-14.
2. I. F. Akyildiz, J. McNair, J. Ho, H. Uzunalioglu and W. Wang, "Mobility management in current and future communications networks", *IEEE Network,* Vol. 12, No. 4, 1998, pp. 39-49.

3. D. Barbara, “Mobile computing and databases - a survey,” *IEEE Transactions on Knowledge and Data Engineering*, Vol. 11, No. 1, 1999, pp. 108-117.
4. C. H. Cai, W. C. Fu, C. H. Cheng and W. W. Kwong, “Mining association rules with weighted items,” *The International Database Engineering and Applications Symposium*, 1998, pp. 68-77.
5. M. S. Chen, J. Han and P. S. Yu, “Data mining: an overview from a database perspective,” *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 6, 1996, pp. 866-883.
6. R. V. J. Chintalapati, V. Kumar and A. Datta, "An adaptive location management algorithm for mobile computing", *The 22th Annual IEEE Conference on Local Computer Networks,* 1997, pp. 133-140.
7. I. Han and D. H. Cho,"Group location management for mobile subscribers on transportation systems in mobile communication networks", *IEEE Transactions on Vehicular Technology*, Vol. 53, No. 1, 2004, pp. 181-191.
8. T. P. Hong, C. S. Kuo and S. C. Chi, "A data mining algorithm for transaction data with quantitative values," *Intelligent Data Analysis,* Vol. 3, No. 5, 1999, pp. 363-376.
9. T. P. Hong, K. Y. Lin and S. L. Wang, "Mining linguistic browsing patterns in the world wide web," Soft Computing, Vol. 6, No. 5, 2002, pp. 329-336.
10. N. E. Kruijt, D. Sparreboom, F. C. Schoute and R. Prasad, "Location management strategies for cellular mobile networks", *IEEE Electronics & Communication Engineering Journal*, Vol. 10, No.2, 1998, pp. 64-72.
11. W. Ma and Y. Fang, “A new location management strategy based on user mobility pattern for wireless networks”, *The 27th Annual IEEE Conference on Local Computer Networks, 2002.*
12. W. C. Peng and M. S. Chen, ”Developing data allocation schemes by incremental mining of user moving patterns in a mobile computing system”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 1, 2003, pp. 70-85.
13. Y. Saygin and O. Ulusoy, “Exploiting data mining techniques for broadcasting data in mobile computing environments,” *IEEE Transactions on Knowledge and Data Engineering*, Vol. 14, No. 6, 2002, pp. 1387-1399.
14. K.Wang, J. M. Liao and J. M Chen, “Intelligent location tracking strategy in PCS”, *The IEE Proceedings on Communications,* 2000, Vol. 147, No. 1, pp. 63-68.
15. S. Yue, E. Tsang, D. Yeung and D. Shi, “Mining fuzzy association rules with weighted items,” *The IEEE International Conference on Systems, Man and Cybernetics*, 2000, pp. 1906-1911.
16. L. A. Zadeh, “Fuzzy logic,” *IEEE Computer*, 1988, pp.83-93.
17. H. J. Zimmermann, *Fuzzy Set Theory and Its Applications*, Kluwer Academic Publisher, Boston, 1991.

Individualized Product Design by Evolutionary Algorithms

Maik Maurer and Udo Lindemann

TU München, Institute of Product Development
Boltzmannstr. 15, 85747 Garching
{maurer,lindemann}@pe.mw.tum.de

Abstract. The steady trend towards product customization requires new strategies in product design. A promising approach aims at the creation of predesigned product structures optimized for later component adaptations. Thus, if customer demands turn up, change impacts are minimized and can be easier handled in order to diminish delivery time and costs. As available methods of structure planning show disadvantages in comprehensive considerations of interdependency networks, an evolutionary algorithm based process has been developed for mediating case specific structural characteristics to the designer. In this context a fitness function has been implemented to meet the selection demands. Further, the new approach permits automated structure optimization by modifying single product dependencies. The designer obtains propositions for useful product adaptation regarding product's aptitude for later customization.

1 The Trend Towards Product Customization

Nowadays, a multitude of manufacturers offers numerous variant models of one basic and formerly standardized product. Reasons for this strategy can be found in partial saturation of markets, more differentiated customer requirements, or the general transition from a supplier-controlled to a purchaser-controlled market paradigm. The increase in variant models induces rising expenses in product design that motivate the use of diverse strategies to control resulting complexity [1]. A modular build-up of parts allows generating a sizeable number of variant models with reasonable effort to meet manifold customers' expectations. Unfortunately, the simple combining of the given number of parts does not forcibly come along with fully individualized products. The core problem in offering a wide number of variant models is that it does not fully accommodate the customer with a product, which is fully suited for his needs. The customer has to choose from a range of goods on a given market.

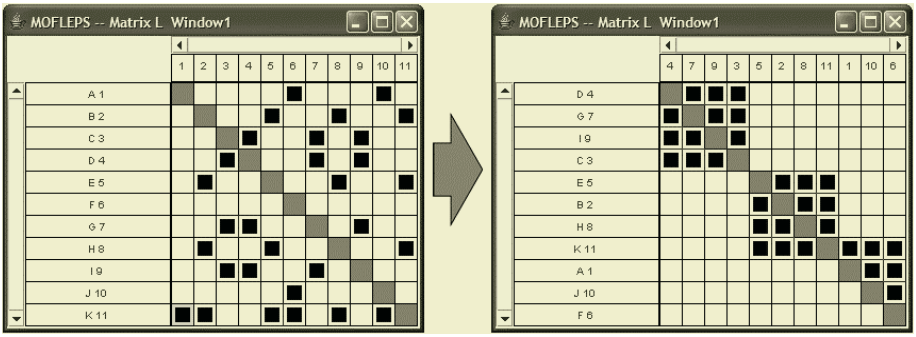
A new approach in product design aims at the creation of product structures, which persist being manageable when been adapted due to (only short-term known) customer demands. This means to handle comprehensive knowledge on change impacts in complex product structures. If the product designer is able to determine directly the elements of a complex network, which have to be adapted, he can indicate expected production time and costs. If he possesses information on an accumulation of customization demands in a specific product range, he can adapt this range in order to minimize possible change impacts to the product. Hereby achieved is a structure optimization concerning forecasted tendencies of customization.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1359–1365, 2005.

2 Enhancement on Conventional Design Planning

The known methodology of variant design [1] is only partly applicable for planning a product adaptable by customers. The immense amount of possible customization requests could only be met if product modularization would take place on a very high level of detail. However, this would lead to unacceptable high development time and costs (definition and design of numerous interfaces) and would result in suboptimal solutions for the specific product functionality.

A promising approach for designing customizable products [2] is based on the analysis of so-called change dependencies. A change dependency located between two elements describes which element will require additional adaptation due to a primarily initiated element adaptation. Methods and tools exist to acquire such networks of dependencies [3, 4, 5] and deterministic algorithms are in use for identification of specific substructures. Designers often use adjacency matrices for interacting with such structures and the proceeding has been systemized in the design structure matrix [3]. Figure 1 shows an exemplarily product structure, where identification of specific substructures corresponds to a re-sorting of elements in the adjacency matrix.



MOFLEPS -- Matrix L Window1

	1	2	3	4	5	6	7	8	9	10	11
A 1	▨					■				■	
B 2		▨			■			■			■
C 3			▨	■			■		■		
D 4			■	▨			■		■		
E 5		■			▨			■			■
F 6						▨					
G 7			■	■			▨		■		
H 8		■			■			▨			■
I 9			■	■			■		▨		
J 10						■				▨	
K 11	■	■			■	■		■		■	▨

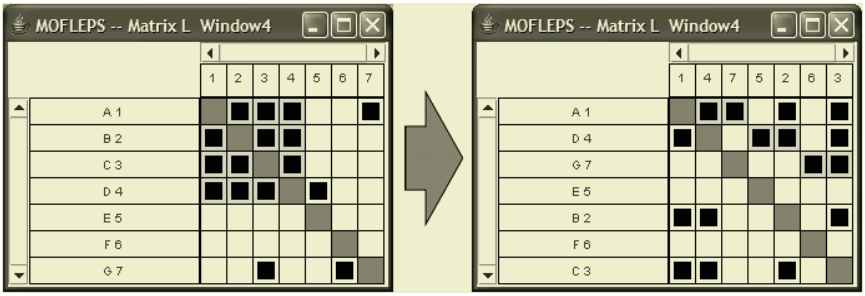
MOFLEPS -- Matrix L Window1

	4	7	9	3	5	2	8	11	1	10	6
D 4	▨	■	■	■							
G 7	■	▨	■	■							
I 9	■	■	▨	■							
C 3	■	■	■	▨							
E 5					▨	■	■	■			
B 2					■	▨	■	■			
H 8					■	■	▨	■			
K 11					■	■	■	▨	■	■	■
A 1									▨	■	■
J 10										▨	■
F 6											▨

Fig. 1. Product structure representation in adjacency matrices

Both matrices represent exactly the same structure, but only the element alignment in the right matrix points out the existence of two complete clusters and a fully connected hierarchy. These substructures are important for understanding the system characteristics, thus designers need information about them to create and optimize customizable products. For example, the knowledge about feedback loops helps to avoid unintentional self-energizing effects, when adapting a single element; hierarchical substructures will complicate customization processes because of chain reactions resulting from single adaptations. In typical change dependency networks structural attributes often superpose each other, exist incompletely, and are of different relevance depending on the use case. The dilemma of structure analysis by deterministic algorithms can be observed in the exemplarily structure of figure 2. Depending on element alignment, a hierarchical or a cluster substructure can be identified. The alignments of structural characteristics influence each other, but multiple structure criteria have to be considered, when comprehensive analyses have to be carried out.

For this reason, a new approach for the comprehensive analysis of change dependency structures is presented here. It bases on an evolutionary algorithm [6] and per-



	1	2	3	4	5	6	7
A 1	▨	■	■	■			■
B 2	■	▨	■	■			
C 3	■	■	▨	■			
D 4	■	■	■	▨	■		
E 5					▨		
F 6						▨	
G 7			■			■	▨

	1	4	7	5	2	6	3
A 1	▨	■	■		■		■
D 4	■	▨		■	■		■
G 7			▨			■	■
E 5				▨			
B 2	■	■			▨		■
F 6						▨	
C 3	■	■			■		▨

Fig. 2. Identification of substructures, depending on matrix alignment

mits the simultaneous consideration [7] of four basic structures: cluster, hierarchy, bus, and single dependency positioning. The approach allows the determination of one matrix alignment, optimizing the comprised interdepending substructure arrangements by case specific relevance. If e.g. the identification of hierarchical dependencies is of major importance, related substructures will be pointed out, also if other existing substructures become invisible in the matrix. Further substructures, which are not affected by hierarchy alignment, will be correctly (visual recognizably) arranged. Thus, the product designer obtains maximum information on structural content with focus on desired criteria. In contrast to known matrix sorting, the new approach also considers and points out incomplete substructures. In addition, the approach serves for an active optimization of product structures by proposing meaningful adaptations as a scenario for further development. Designers can set a maximum number of dependency modifications and receive an optimized structure as well as information on executed modifications. Alternatively, one can influence the desired structural result (e.g. "no feedback loops between the elements A and B") and obtains propositions on required modifications.

3 Basic Design of the Evolutionary Algorithm Approach

To meet the requirements explained above, the presented approach uses an approved evolutionary algorithm [8] and integrates a specially designed fitness function as well as a preliminary analysis phase. The process layout is shown in figure 3. First, the structure in question is analyzed to identify specifications of basic substructures. This information about comprised specifications of clusters, hierarchies, busses, and single dependency positioning is required for later fitness determination, where it serves as calculation base for the relative quantification of structure alignments. In common, designers are familiar with cluster, bus, and hierarchy structures and their alignment in adjacency matrices [3, 9]. The additionally considered single dependency positioning describes the location of dependencies in the matrix, which do not belong to further basic structures. Best arrangement for these single dependencies is (due to easier matrix reading) close to the diagonal. The identification is separately performed for every basic structure, without consideration of mutual impacts. For this reason, deterministic algorithms of the graph theory are applied for the analysis process.

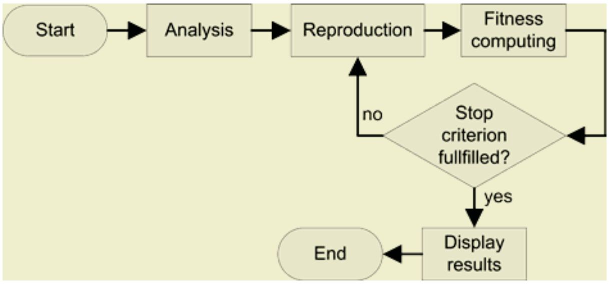


Fig. 3. Basic process of the evolutionary algorithm approach

If the existing quantity and specification of basic structures is identified, the reproduction process is initiated. As effective reproduction procedures are available, an appropriate one has been directly taken from [8] without major adaptations. Even if the reproduction worked well in later evaluation, specific design of reproduction offers possibilities of process optimization in the future. In contrast to the reproduction, the fitness function has been newly designed in order to meet the existing requirements. The implementation is explained in the following chapter.

4 The Fitness Function

The calculated fitness value of a matrix alignment results from the comparison of actual alignment to ideal alignment of each basic structure specification determined in the analysis step [6, 7]. Because of mutual impact between basic structure alignments (see figure 2), some matrices cannot reach a previously analyzed ideal alignment. Thus, it is required to specify the importance of particular basic structures by weighting factors in order to provide a decision support in case of conflicts.

The fitness value for bus structures and single dependencies results directly from their position in the adjacency matrix. To determine the fitness values for clusters and hierarchy structures, their ideal alignment is required. The operation sequence of fitness calculation is shown in figure 4. For every basic structure, the fitness value of its identified specifications is sequentially computed and finally added to the entire fitness value of the matrix alignment.

Applied criteria for evaluating the fitness of basic structure specifications are described in the following: the fitness value of bus structures depends on its absolute position in the matrix; busses of incoming and busses of outgoing edges are considered separately. It must be taken into account that if more than one bus (of the same type of edge) exists in the matrix, they have to be ranked before assigning appropriate positions. Busses are best arranged at the left or right borders of the matrix.

When determining cluster fitness values, completely and incompletely interconnected clusters are considered separately, as element alignment in-between the cluster is irrelevant for previous ones (see fig. 1). In incompletely interconnected clusters, the fitness value increases with the quantity of edges located nearby the matrix diagonal. A fitness criterion for both cluster types is the degree of connected alignment of belonging cluster elements. If less than 70% of the cluster's elements are aligned connectedly, the fitness value is zero, as visual recognition of the cluster is not possible.

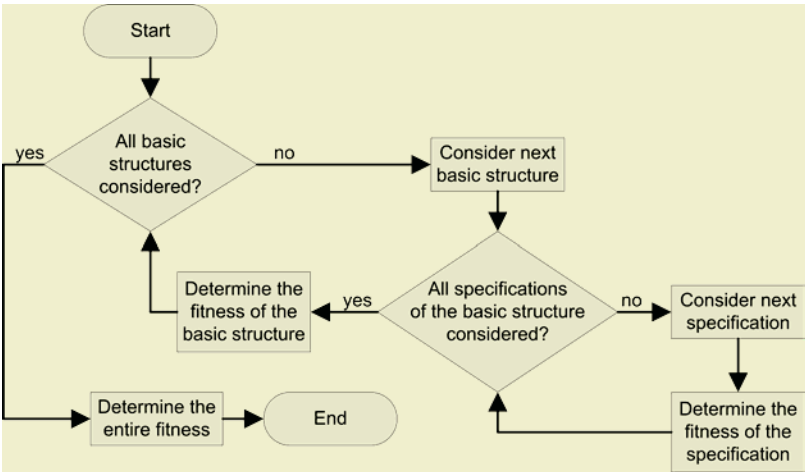

Start
All basic structures considered?
yes
no
Consider next basic structure
All specifications of the basic structure considered?
yes
no
Determine the fitness of the basic structure
Consider next specification
Determine the fitness of the specification
Determine the entire fitness
End

Fig. 4. Process of fitness determination

Fitness values for hierarchical structures increase with adequate (visually recognizable) alignment of nodes in hierarchies. In contrast to cluster fitness, it is not the proximity to the diagonal, but the distinct perceptibility of hierarchy levels. Second criterion for hierarchy examination is the degree of connected alignment of implied nodes.

Fitness values for single dependencies (which are not integral part of other basic structure's specifications) depend on the dependency's vertical distance from the matrix diagonal. The closer it gets the higher fitness value results. In addition, the quantity of existing single dependencies as well as the positioning of other basic structure's specifications must be taken into account, as the quantity of single dependencies may exceed the available positions close to the matrix diagonal.

5 Enhancement Towards Structure Optimization

By use of the shown approach, product structures can be efficiently identified in matrices by predetermined criteria. An aspect for further development of product structures is the automated structure optimization by modification of specific dependencies. Small changes in the dependency network often cause major structural consequences already. For example, the removal of a high-level dependency in a hierarchy structure will clearly diminish this hierarchy. If designers aim at simplified structures with less change dependencies (e.g. when planning customizable products), the software implementation will automatically propose such adaptations. Thus, the designer obtains promising propositions for optimization he can verify for practical feasibility. In contrast to structure analyses, now the product structure is changed with the objective to generate a best possible structure arrangement - by as less modifications as required. The enhancement to the process of structure analysis is shown in figure 5.

Newly introduced is the step of dependency adaptation, which is executed prior to the matrix reproduction. Another analysis step is required afterwards, as by modifying the matrix content the determination base for later fitness evaluation might change. A second stop criterion is implemented, which manages the decision on further modification after optimizing a matrix alignment. Thus, a re-alignment process follows every structure modification in order to provide perceivable matrix visualization.

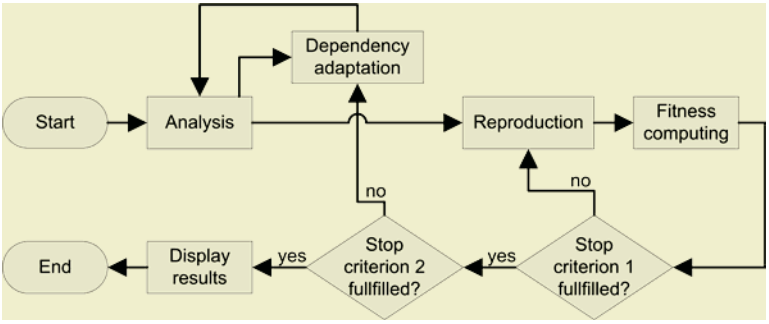


Fig. 5. Evolutionary algorithm enhanced for structure adaptation

6 Verification

The presented approach of structure analysis has been evaluated by 160 adjacency matrices containing 10 elements each. The low sized test matrices were well suited for manual revision of automatically optimized alignment. Further research was examined with the 25 elements containing dependency network of an industrial packing machine. The two matrices in figure 6 show the results of the structure analysis without dependency adaptation (left matrix) and after modification of four dependencies (right matrix). The optimized alignment shows three clusters overlapping each other. Their internal degree of interconnectivity is relatively low and a large amount of feedback loops could be identified. Thus, probable dependency modifications were required in order to simplify the structure and therefore enable an easier product handling. By removal of four dependencies, the clusters were optimized and feedback loops were reduced to 60%. These modifications propositions provided useful support for the further product design.

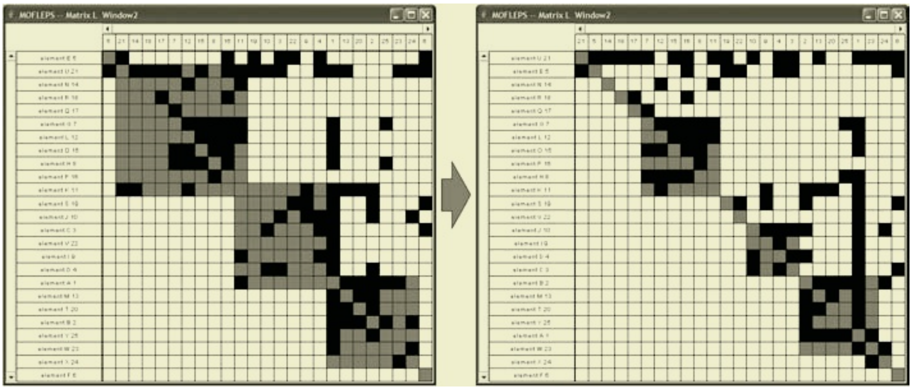


Fig. 6. Proposed structure modifications for product optimization

The presented approach supports designers in structure identification as well as modification by multiple structural criteria. It helps visualizing substructures relevant for the specific use case. Modification propositions can be derived, enabling designers to focus on decisive product parts. The choice of convenient weighting in fitness determination is important for the resulting structure quality. In further work, suitable weighting factors for common use cases will be acquired.

References

1. Ericsson A.;Erixon G.: Controlling Design Variants – Modular Product Platforms. ASME Press: New York 1999.
2. Maurer, M.; Pulm, U.; Lindemann, U.: A process model for designing fuzzy product structures. In: Marjanovic, D. (Ed.): 8th International Design Conference – DESIGN 2004, Dubrovnik (Croatia). Zagreb: Sveucilisna tiskara 2004. pp. 383-388.
3. Browning T. R.: Applying the Design Structure Matrix to System Decomposition and Integration Problems: A Review and New Directions. In: IEEE Transactions on Engineering Management, 2001.
4. Rogers, J. L.: DeMAID/GA: An Enhanced Design Manager's Aid for Intelligent Decomposition. Langley: NASA Langley Research Center, 1996.
5. Maurer, M.; Pulm, U.; Lindemann, U.: Utilization of graph constellations for the development of customizable product spectra. In: Fourth International ICSC Symposium on Engineering of Intelligent Systems EIS 2004, Madeira: ICSC Interdisciplinary Research Canada 2004.
6. Yu, T.-L.; Goldberg, D., E.; Yassine, A.; Chen, Y.-P.: A Genetic Algorithm Design Inspired by Organizational Theory: A Pilot Study of a Dependency Structure Driven Genetic Algorithm. IlliGAL Report No. 2003007, February 2003a.
7. Yu, T.-L.; Yassine, A; Goldberg, D., E.: A Genetic Algorithm for Developing Modular Product Architectures. IlliGAL Report No. 2003024, October 2003b.
8. Goldberg, D. E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Reading: Addison-Wesley 1989.
9. Gibbons, A.: Algorithmic graph theory. Cambridge: Cambridge Univ. Press 1999.

Fuzzy Similarity Measure and Fractional Image Query for Large Scale Protein 2D Gel Electrophoresis[*]

Daw-Tung Lin[1], Juin-Lin Kuo[2], En-Chung Lin[3], and San-Yuan Huang[4]

[1] Department of Computer Science and Information Engineering
National Taipei University, 151, University Rd., San Shia, Taipei, 237 Taiwan
[2] Biomedical Engineering Center, ITRI
Hsinchu, Taiwan
[3] Department of Animal Science
National Taiwan University, Taipei, Taiwan
[4] Divisions of Applied Biology and Biotechnology
Animal Technology Institute Taiwan, P.O. BOX 23, Chunan, Miaoli, Taiwan

Abstract. Spot matching is a challenging problem in two dimensional protein gel electrophoresis (2DGE) chromatography images analysis. It is necessary to provide a robust solution to the problem of querying and matching large scale for various size of 2DGE images. In this paper, we have developed a novel maximum relation spanning tree (MRST) algorithm which is capable of performing fast and accurate matching without the need for landmarks to be manually selected. In the matching process, we employ fuzzy similarity measuring technique to conclude the final decision of matching and location. The resultant system performs up to 94% correct matching for 225 2DGE test images. The additive value is the foundation of querying fractional gel images with large format gel images database.

1 Introduction

In the research of protein expression analysis, two-dimensional gel electrophoresis (2DGE) chromatography is a popular tool for investigating differential patterns of qualitative protein expression [1]. The problems can be categorized into image registration, image distortion correction, spot detection, and spot matching. Two dimensional spot matching of two non-uniform images is an NP-hard problem [2]. Their computation is non-deterministic polynomial time. A few algorithms have been proposed and tried to solve this problem, for example Restriction Landmark Genomic Scanning (RLGS) [3–5] and Fuzzy Cluster [6]. RLGS compares the protein using construction of computer graphs and landmark. Fuzzy Cluster method uses the relation between two protein spots and calculates the similarity. For two gel images with n and m spots, the worst case upper bound of computation complexity is $O(n^2m^2)$ arc pairs and $O(n \log m)$ for measuring the pattern

[*] This work was supported in part by the National Science Council, Taiwan, R.O.C. grants NSC90-2313-B-059-002, NSC91-2745-E-216-001, and NSC93-2213-E-216-016

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1366–1372, 2005.

similarity of each pair [7]. In this paper, we propose a novel Maximum Relation Spanning Tree (MRST) and integrate fuzzy inference technique to solve the matching problem. We can use this method to find the gel image which contains or is similar to the small and fractional query image, and to locate the area residing in the large scale images. In addition, this method is fully automated and does not need landmark allocated in a priori by users.

2 Features Extraction and Fuzzy Similarity Measure

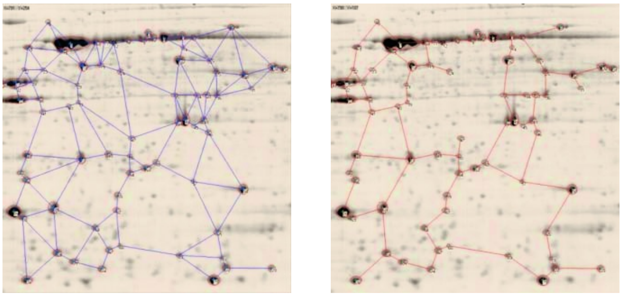
We have to construct features of the protein spots which are invariant to intensity bias and geometric distortions due to the casting, polymerization and running procedure of the gels [7]. In this research, we apply the computer graphics theory to construct and to extract the features from 2DGE images [8–10]. We have selected the Gabriel Graph (GG) [11] and the Relative Neighborhood Graph (RNG) [3] as the feature construction models because the variation of point's feature is more obvious than that of the others. The Gabriel graph P , denoted by $GG(P)$, has its region of influence over the closed disk having segment $\overline{uv}$ as diameter. That is, two vertices $u, v \in S$ are adjacent if and only if

$$D^2(u,v) < D^2(u,w) + D^2(v,w),\ \ for\ all\ w \in V,\ \ w \neq u,\ \ v. \tag{1}$$

where $D(u, v)$ denotes the distance of $\overline{uv}$. The relative neighborhood graph of P, denoted by $RNG(P)$, has a segment between points u and v in P if the intersection of the open disks of radius $D(u, v)$ centered at u and v is empty. Equivalently, $u, v \in V$ are adjacent if and only if

$$D(u,v) \leq max[D(u,w), D(v,w)], for\ all\ w \in V, w \neq u, v. \tag{2}$$

Thus, RNG is a subset of a GG and is relatively transformation insensitive compared with its superset [3]. Examples of Gabriel graph and Relative neighborhood graph constructed from one gel image are shown in Fig 1(a) and (b), respectively. Geometrical spot matching relies on the similarity of the features extracted from the structured graphs. After we have constructed the proximity graphs, we continue to extract the features of the spots. For each node on both



(a) Gabriel graph (b) Relative neighborhood graph

Fig. 1. Examples of graph representation for a 2D gel image

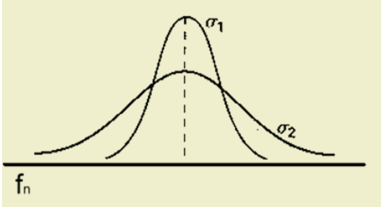


Fig. 2. Illustration of membership function

Gabriel Graph and Relative Neighborhood Graph, three features are obtained: f_1 degree of each protein spot, f_2 angle of connected edges, and f_3 Euclidean distance between protein spots. The conventional direct superimpose matching is not appropriate due to imperfect 2D electrophoresis technique [7]. An adaptive decision method is essential to examine and measure the similarity from the above-mentioned spot pair features. We decide to apply the fuzzy inference to develop our comparative framework. Due to the difference in each local area of the 2DGE spots geometric relation, we use a stylized gaussian membership function as shown in Fig. 2. The nth feature of spot s in the sample gel image is defined as $f_n(s)$, and the corresponding feature of spot r in the reference gel image is defined as $f_n(r)$. Let μ_{f_n} be the fuzzy membership function of the similarity measure between the sample image and reference image according to the nth feature:

$$\mu_{f_n}(s,r) = e^{-\frac{(f_n(s)-f_n(r))^2}{2\sigma^2}}, \tag{3}$$

where σ denotes the variance of the feature f_n between spots. The function is illustrated in Fig 2 in which σ_1 and σ_2 denote different local intensity. In this figure, we can see that different sets of spots will have different kind of membership functions constructed by different σ. With three different features, we calculate three fuzzy relations $\mu_{f_1}, \mu_{f_2}, \mu_{f_3}$ for distance, degree, and angle on the Garbriel graph, respectively. To aggregate three fuzzy measurements, a weighted mean value [12] is computed and defined as the closure measurement:

$$R(s,r,\omega_{f_1},\omega_{f_2},\omega_{f_3}) = \frac{\omega_{f_1}\cdot\mu_{f_1}(s,r)+\omega_{f_2}\cdot\mu_{f_2}(s,r)+\omega_{f_1}\cdot\mu_{f_3}(s,r)}{3}, \tag{4}$$

where w_{f_i} is the weight of the corresponding feature and $\Sigma_{i=1}^{3}\omega_{f_n} = 3$. The weights can be set optimally according to learning mechanism. Finally, we can choose the maximum relationship from the spot pairs and proceed to the next comparison procedure.

3 Fractional Matching with Maximum Relation Spanning Tree

In order to compare the similarity between two gel images, we have developed a maximum relation spanning tree (MRST) algorithm, in which the minimum distance derived from the Minimum-Cost Spanning Tree [13] is replaced by the proposed maximum fuzzy relation. We calculate the relationship between protein points using the features of the structured graphs and find the maximum relationship of fuzzy inference. The MRST algorithm is elucidated as follows. When we implement this algorithm, we separate the process into two parts: global matching and local matching.

```
Maximum Relation Spanning Tree Algorithm
MRST() {
 If node tree T is empty
   Insert a new anchor point pair with the maximum fuzzy relation;
   MRST();
 else if  anchor point pairs are not empty
          Find next anchor point pair in the satellite spots;
          MRST();
      else terminate;
 Compare the matched area;
 If the difference is less than 10%
   then it is matched
   else match fail;
}
```

3.1 Global Matching

In this step we find the initial anchor point pair automatically by comparing the fuzzy relation of all possible corresponding anchor pairs between the sample gel image and reference gel image. Substituting the feature of degree from the profiles of GG and RNG into the similarity measure of fuzzy membership function (Equation 3), we can obtain two fuzzy relationship values R_{GG} and R_{RNG}, respectively. If the product $R_{GG} \times R_{RNG} \geq 0.7$, then we treat it as a candidate anchor point pairs with larger fuzzy similarity relationship. The candidate pair with the maximum fuzzy relation will be selected with higher priority in the local matching process.

3.2 Local Matching

Once the anchor point is located, we start to apply the maximum relation spanning tree algorithm on the Gabriel graph. The fuzzy similarity measure (Equation 4) of three features (degree, distance, and angle) of the Gabriel graph is computed. If the fuzzy membership is greater than 0.6, the graph is extended. This algorithm will proceed recursively until all the spot pairs produced by Gabriel matching is completed. The flow chart is depicted in Fig. 3. Through this process, we will find all similar spot pairs between two gel images.

4 Simulation Results

We have implemented and tested the proposed system. The experiments were based on fifteen 2D protein gel profiles (image size: 1498 x 1544) of porcine

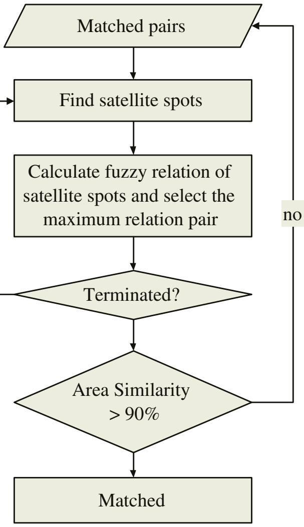

Matched pairs
Find satellite spots
Calculate fuzzy relation of satellite spots and select the maximum relation pair
no
no
Terminated?
Area Similarity > 90%
Matched

Fig. 3. The flowchart local matching of the protein 2D gel pattern matching process

testis obtained from the Bioinformatics Center for Swine Research in Tropical Area at the Animal Technology Institute Taiwan (ATIT). The experiment data set is constructed from these fifteen gel images. The data set contains totally 225 gel images as follows: 15 original gel images (1498×1544), 135 fractional gel images composed of nine different sizes of fractional images chopped randomly from each of 15 original gel images (listed in Table 1), and 75 rotated images: 45 gel images obtained from the original gel images by rotating in 90°, 180°, and 270° degrees, respectively, and 30 gel images obtained from the original gel images by flipping horizontally and vertically, respectively. We have done the test on fractional matching by using various size of segmented image samples (135 images) to perform query in the original gel images (15 images). The correct match is up to 94%. The results are detailed in Table 1. In order to simulate the situations of image rotation, reverse, and translation, we have tested 75 different modified gel images with five situations mentioned above. The ratio of correct matching is 100%. To further confirm the capability of fractional matching, we have also used the rotated fractional gel images and to perform searching in the original large-scale gel images. One of the results is demonstrated in Fig. 4 where one fractional sample gel image of size of 200×200 is rotated or flipped into five images with different conditions (rotated in 90°, 180°, 270°, flipped horizontally ,and vertically) shown on the left hand side in Fig. 4. With these six small images, we tried to search in the original large 2D gel image database. The location of correct matching is identified in the rectangle on the right-hand side in Fig. 4. Only Pánek and Vohradský reported their matching accuracy is 98% but based on one gel image [14]. To justify the advantage of the proposed work, we need to make comparison with other methods. However, the quantitative information of fractional matching performance is not available from the literature survey.

Table 1. The results of fractional matching of different size of images with adapted parameters

	Correct Matching Ratio
Original Images	100 %
1000x1000	100 %
900x900	100 %
800x800	100 %
700x700	100 %
600x600	100 %
500x500	100 %
400x400	100 %
300x300	86.7 %
200x200	53.3 %
Overall	94 %

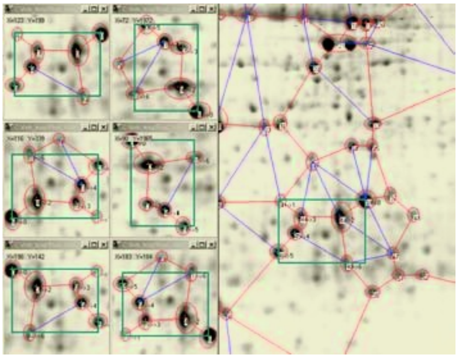


Fig. 4. Result of fractional matching and allocation processing

5 Conclusion

In the research of differential patterns investigation for qualitative protein expression, it is necessary to provide a robust solution to the problem of querying and matching large scale and large sets of protein 2DGE chromatography. In this paper, we have developed a novel, fast, accurate and content-based image matching method MRST utilizing the fuzzy inference technique. We have selected the Gabriel Graph and Relative Neighborhood Graph as the feature construction models. It is expected to compensate the variance of geometric distortions automatically. The proposed method not only can handle the rotation, shift and reverse condition, but can also handle fractional mapping problem. We can use this method to find the gel image which contains or is similar to the small and fractional query image, and to locate the area residing in the large scale images. After all, we can constitute the gel images and protein spots information into the database for further investigation. The proposed system achieves up to 94% correct matching in large-scale gel image searching scenarios. Most importantly, the proposed MRST matching algorithm requires neither the landmarks manually set nor a prior information of gel image alignment.

References

1. M. J. Dunn S. Veeser and G. Z. Yang. Multiresolution image registration for two-dimensional gel electrophoresis. *Proteomics*, 1:856–870, 2001.
2. Tatsuya Akutsu, Kyotetsu Kanaya, Akira Ohyama, and Asao Fujiyama. Point matching under non-uniform distortions. *Discrete Appl. Math.*, 127(1), 2003.
3. Y. Watanabe K. Takahashi, M. Nakazawa and A. Konagaya. Fully-automated spot recognition and matching algorithms for 2-D gel electrophoretogram of genomic DNA. In *Genome Informatics Workshop*, pages 161–172, 1998.
4. K. Takahashi, Y. Watanabe M. Nakazawa, and A. Konagaya. Automated processing of 2-D gel electrophoretograms of genomic DNA for hunting pathogenic DNA melecular changes. In *Genome Informatics Workshop*, pages 121–132, 1999.
5. T. Matsuyama, T. Abe, C. H. Bae, Y. Takahashi, R. Kiuchi, T. Nakano, T. Asami, and S. Yoshida. Adaptation of restriction landmark genomic scanning to plant genome analysis. *Plant Molecular Biology Reporter*, 18:331–338, 2000.
6. X. Ye, C.Y. Suen, and E. Wang. M. Cheriet. A recent development in image analysis of electrophoresis gels. In *Vision Interface (VI'99), Trois-Rivieres, CA, 19-21*, pages 432–438, 1999.
7. A.W. Dowsey, M.J. Dunn, and G.Z. Yang. The role of bioinformatics in two-dimensional gel electrophoresis. *Proteomics.*, 3:1567–1596, 2003.
8. A. Efrat, F. Hoffmann, K. Kriegel, and C. Schultz. Geometric algorithms for the analysis of 2D-electrophoresis gels. *Journal of Computational Biology*, 9(2):299–315, 2002.
9. F. Hoffmann, K. Kriegel, and C. Wenk. Matching 2D patterns of protein spots. In *Symposium on Computational Geometry 1998*, pages 231–239, 1998.
10. F. Hoffmann, K. Kriegel, and C. Wenk. An applied point pattern matching problem: comparing 2D patterns of protein spots. In *Discrete Applied Mathematics 93*, pages 75–88, 1999.
11. J. I. Garrels. Two-dimensional gel electrophoresis and computer analysis of proteins synthesized by clonal cell lines. *Biological Chemistry*, 254:796l–7977, 1979.
12. G. J. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic - Theory and Appication.* Prentice Hall International Editions, 1995.
13. E. Horowitz, S. Sahni, and D. Mehta. *Fundamentals of Data Structures in C++.* Computer Science Press, 1995.
14. J. Pánek and J. Vohradský. Point pattern matching in the analysis of two-dimensional gel electropherograms. *Electrophoresis*, 20:3483–3491, 1999.

Building the Fuzzy Control System Based on the Pilot Knowledge

Michał Lower[1], Dariusz Król[2], and Bogusław Szlachetko[3]

[1] Institute of Cybernetics
Wroclaw University of Technology, Poland
[2] Institute of Applied Informatics
Wroclaw University of Technology, Poland
[3] Institute of Telecommunication and Acoustics
Wroclaw University of Technology, Poland

Abstract. The main problem addressed in this paper is simplification of transforming the knowledge from natural language description to fuzzy rules control system. The proposed fuzzy system build on the observation of human pilot reaction and on few very simple rules is very promising. Our fuzzy control system allow to control the helicopter in the hover yet can be generalize to other aspect of helicopter flight.

1 Introduction

There exists the well-founded assumption, that the build of the simple rules fuzzy control system for taking control on helicopter in hover state is possible. The human makes decision basing on few rules and round-off variables [1, 4, 5, 9]. Human mind use the indeterminate abstraction which are hard to measure. Nevertheless human pilot can control much complicated object like helicopter. Even the beginner pilot can control helicopter in hover state on first training stage.

Therefor build of the fuzzy control system which can automatically control helicopter in hover is possible. This fuzzy system will be able to function like the beginner pilot.

2 Linguistic Description of the Helicopter System

The first stage of this fuzzy control system build is creating of the linguistics strategy of control basing on the qualitative describe of the helicopter. The main source of information, in this stage, are: the didactic materials of helicopter pilotage course [11], observation of the fuselage reaction on the steering control and helicopter pilots relation and knowledge. Observation was leading in time of fly with different experience pilots. One of them has fifteen years of practice.

On the base of the preceding materials the following observation was made:

1. Pilot actions on the appropriate control stick has the impulse nature. Pilot inclines control stick for the short time, next opposes it and when occasion

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1373–1379, 2005.

arrives he repeat action decreasing the amplitude of inclines to achieve expected result. Number of inclines, its amplitude and time depend on the pilot experience. The pilot with five years practice does many short impulses when steering actions of the fifteen years practice pilot are good-tempered. More experienced pilot needs only few inclines.

2. In simplification, amplitude of sticks inclination is proportional to largeness of deviation from desired value, whereas duration of inclination and its maximal value depends on speed, direction and dynamic changes of deviation. Typical duration of sticks inclination is bellow 1[s], whereas scope of stick inclination often exceeds 60% of whole inclination range.
3. Every fuselage reaction on the steering along one direction are coupled with other directions. Nevertheless main direction can be infer.
4. Pilot reaction resist mainly on subjective valuation of eyes stimulus and subjective valuation of dynamic position changes register by the ears labyrinth. Indications of deck equipments have only correcting character. Uninterrupted analysis of all indicators of many fly parameters and state of many helicopter systems showed by deck equipment is not possible by human pilot.

3 Linguistic Description to Fuzzy Rules Transformation

The following variables are used to transform the linguistic description of helicopter movement to fuzzy rules set:

$\dot{\Theta}_0$ – a collective pitch angle of blades;
$\dot{\Theta}_1$ – a longitudinal cyclic pitch angle;
$\dot{\Theta}_2$ – a lateral cyclic pitch angle of blades;
$\dot{\Theta}_{so}$ – a collective pitch angle of the tail rotor;
U, V, W – linear velocities along O_x, O_y, O_z respectively
P, Q, R – angular velocities respect of O_x, O_y, O_z axes
Θ, Φ - roll and pitch angles

3.1 Linguistic Fuzzy Rules

Let's define the following linguistic variables:

- χ – which denotes the slow changes of parameters
- Υ – which denotes the normal changes of parameters

Each of linguistic variable can be $(.)^+$ – direction right, up, etc. or $(.)^-$ – direction left, down, etc. Thus we can define the following rules.

Linguistic rules for controlling collective pitch angle of the tail rotor – eq.(1):

- **IF** the angular velocity R is positive $(\chi^+ \ \Upsilon^+)$ **THEN** set the collective pitch angle $\dot{\Theta}_{so}$ to positive $(\chi^+ \ \Upsilon^+)$
- **IF** the linear velocity R is negative $(\chi^+ \ \Upsilon^+)$ **THEN** set the collective pitch angle $\dot{\Theta}_{so}$ to negative $(\chi^+ \ \Upsilon^+)$

Linguistic rules for controlling collective pitch angle – eq.(2):

- **IF** the linear velocity W is positive (χ^+ $\varUpsilon^+$) **THEN** set the collective pitch angle $\dot{\varTheta}_0$ to positive (χ^+ $\varUpsilon^+$)
- **IF** the linear velocity W is negative (χ^+ $\varUpsilon^+$) **THEN** set the collective pitch angle $\dot{\varTheta}_0$ to negative (χ^+ $\varUpsilon^+$)

Linguistic rules for controlling longitudinal cyclic pitch angle – eq.(3):

- **IF** the linear velocity U is positive (χ^+ $\varUpsilon^+$) **AND** angular velocity Q is negative (χ^- $\varUpsilon^-$) **THEN** set the longitudinal cyclic pitch angle $\dot{\varTheta}_1$ to positive (χ^+ $\varUpsilon^+$)
- **IF** the linear velocity U is positive (χ^+ $\varUpsilon^+$) **AND** angular velocity Q is positive (χ^- $\varUpsilon^-$) **THEN** set the longitudinal cyclic pitch angle $\dot{\varTheta}_1$ to negative (χ^+ $\varUpsilon^+$)
- **IF** the linear velocity U is negative (χ^+ $\varUpsilon^+$) **AND** angular velocity Q is position (χ^- $\varUpsilon^-$) **THEN** set the longitudinal cyclic pitch angle $\dot{\varTheta}_1$ to negative (χ^+ $\varUpsilon^+$)
- **IF** the linear velocity U is negative (χ^+ $\varUpsilon^+$) **AND** angular velocity Q is negative (χ^- $\varUpsilon^-$) **THEN** set the longitudinal cyclic pitch angle $\dot{\varTheta}_1$ to positive (χ^+ $\varUpsilon^+$)

Linguistic rules for controlling lateral cyclic pitch angle – eq.(4):

- **IF** the linear velocity V is positive (χ^+ $\varUpsilon^+$) **AND** angular velocity P is negative (χ^- $\varUpsilon^-$) **THEN** set the lateral cyclic pitch angle $\dot{\varTheta}_2$ to positive (χ^+ $\varUpsilon^+$)
- **IF** the linear velocity V is positive (χ^+ $\varUpsilon^+$) **AND** angular velocity P is positive (χ^- $\varUpsilon^-$) **THEN** set the lateral cyclic pitch angle $\dot{\varTheta}_2$ to negative (χ^+ $\varUpsilon^+$)
- **IF** the linear velocity V is negative (χ^+ $\varUpsilon^+$) **AND** angular velocity P is position (χ^- $\varUpsilon^-$) **THEN** set the lateral cyclic pitch angle $\dot{\varTheta}_2$ to negative (χ^+ $\varUpsilon^+$)
- **IF** the linear velocity V is negative (χ^+ $\varUpsilon^+$) **AND** angular velocity P is negative (χ^- $\varUpsilon^-$) **THEN** set the lateral cyclic pitch angle $\dot{\varTheta}_2$ to positive (χ^+ $\varUpsilon^+$)

The above linguistic description of control rules can be directly transformed to fuzzy rules as follows:

$$\begin{aligned}
(R == \varUpsilon^-) &\Rightarrow (\varTheta_{so} = \varUpsilon^-)\\
(R == \varUpsilon^+) &\Rightarrow (\varTheta_{so} = \varUpsilon^+)\\
(R == \chi^-) &\Rightarrow (\varTheta_{so} = \chi^-)\\
(R == \chi^+) &\Rightarrow (\varTheta_{so} = \chi^+)
\end{aligned} \tag{1}$$

$$\begin{aligned}
(W == \varUpsilon^-) &\Rightarrow (\varTheta_0 = \varUpsilon^-)\\
(W == \varUpsilon^+) &\Rightarrow (\varTheta_0 = \varUpsilon^+)\\
(W == \chi^-) &\Rightarrow (\varTheta_0 = \chi^-)\\
(W == \chi^+) &\Rightarrow (\varTheta_0 = \chi^+)
\end{aligned} \tag{2}$$

$$\begin{aligned}
(U == \Upsilon^{+}) \cap (Q == \Upsilon^{-}) &\Rightarrow (\Theta_1 = \Upsilon^{+})\\
(U == \Upsilon^{+}) \cap (Q == \Upsilon^{+}) &\Rightarrow (\Theta_1 = \Upsilon^{-})\\
(U == \Upsilon^{-}) \cap (Q == \Upsilon^{+}) &\Rightarrow (\Theta_1 = \Upsilon^{-})\\
(U == \Upsilon^{-}) \cap (Q == \Upsilon^{-}) &\Rightarrow (\Theta_1 = \Upsilon^{+})\\
(U == \chi^{+}) \cap (Q == \chi^{-}) &\Rightarrow (\Theta_1 = \chi^{+})\\
(U == \chi^{+}) \cap (Q == \chi^{+}) &\Rightarrow (\Theta_1 = \chi^{-})\\
(U == \chi^{-}) \cap (Q == \chi^{+}) &\Rightarrow (\Theta_1 = \chi^{-})\\
(U == \chi^{-}) \cap (Q == \chi^{-}) &\Rightarrow (\Theta_1 = \chi^{+})
\end{aligned} \tag{3}$$

$$\begin{aligned}
(V == \Upsilon^{+}) \cap (P == \Upsilon^{+}) &\Rightarrow (\Theta_2 = \Upsilon^{+})\\
(V == \Upsilon^{+}) \cap (P == \Upsilon^{-}) &\Rightarrow (\Theta_2 = \Upsilon^{-})\\
(V == \Upsilon^{-}) \cap (P == \Upsilon^{-}) &\Rightarrow (\Theta_2 = \Upsilon^{-})\\
(V == \Upsilon^{-}) \cap (P == \Upsilon^{+}) &\Rightarrow (\Theta_2 = \Upsilon^{+})\\
(V == \chi^{+}) \cap (P == \chi^{+}) &\Rightarrow (\Theta_2 = \chi^{+})\\
(V == \chi^{+}) \cap (P == \chi^{-}) &\Rightarrow (\Theta_2 = \chi^{-})\\
(V == \chi^{-}) \cap (P == \chi^{-}) &\Rightarrow (\Theta_2 = \chi^{-})\\
(V == \chi^{-}) \cap (P == \chi^{+}) &\Rightarrow (\Theta_2 = \chi^{+})
\end{aligned} \tag{4}$$

3.2 Qualitative Description of Membership Functions

For the simplicity of the model triangular membership function was used. Values of the linguistics variables are presented on the Tables 1.

Table 1. Qualitative values of the angles

Parameter $[rad/s]$	Υ^{-}	χ^{-}	χ^{+}	Υ^{+}
Θ_0	-0.05	-0.03	0.03	0.05
Θ_1	-0.14	-0.10	0.10	0.14
Θ_2	-0.16	-0.12	0.12	0.16
Θ_{so}	-0.26	-0.19	0.19	0.26

For the qualitative description of the flight parameters the helicopter model response on the steering impulse (look at the Table 1) was measured. Each impulse has the duration time equals $1[s]$. Flight parameters trends quantify was the goal. Detailed description of this experiments was done in [9], but in effect linear proportion of the steering impulse to the corresponding parameters trend was observed.

On the Table 2 the qualitative values of the parameters trends used in membership functions are presented. This values are proportional with the constant

Table 2. Qualitative values of the flight parameters

Linguistic values	$U[m/s]$	$V[m/s]$	$W[m/s]$	$P[rad/s]$	$Q[rad/s]$	$R[rad/s]$
Υ^-	2.032	2.583	3.184	3.208	-0.746	-1.996
χ^-	1.452	1.937	1.910	2.406	-0.533	-1.459
χ^+	-1.452	-1.935	-1.910	-2.405	0.533	1.460
Υ^+	-2.023	-2.579	-3.184	-3.207	0.704	1.998

coefficient k, which are heuristically matched, to the measured values. More detail can be find in [9].

4 Simulation – Fuzzy Regulator

Derivatives of the helicopter's steering angles $\mathbf{B} = [\dot{\Theta}_0, \dot{\Theta}_1, \dot{\Theta}_2, \dot{\Theta}_{so}]$ are the input parameters to the simulation model. Output parameters of the simulation model, like in the mathematical model, are the helicopter's longitudinal and angular velocities and the roll and pitch angles. Thus the parameters of the flight vector are: $\mathbf{A} = [U, V, W, P, Q, R]$.

These parameters have been obtained experimentally in PZL Świdnik factory and correspond to the Kania type helicopters.

In simulation the input values of the helicopter model were chosen from the range as follows:

- $\Theta_0 \in (1°; 13°)$, the range of collective pitch angle equals 12°,
- $\Theta_1 \in (-7°; 6°)$, the range of lateral cyclic pitch angle equals 13°,
- $\Theta_2 \in (-5°; 5°)$, the range of longitudinal cyclic pitch angle equals 10°,
- $\Theta_{so} \in (-10°; 20°)$, the range of tail rotor collective pitch angle equals 30°.

Definition of the disturbances induced by the wind can be find in the literature [9, 10]. Influence of this type of disturbances was tested separately for every linear velocity in the system of coordinates $Oxyz$ connected with fuselage. In simulations the restrained blow was used.

5 Fuzzy Control System

Base on the mathematical model of helicopter defined in [2, 3, 7] the simulation model presented in Figure 1 was developed. As the effect the Simmulink application was build with four regulators as follows:

$$\begin{aligned} \dot{\Theta}_0 &= \mathrm{FIS1}(W) \\ \dot{\Theta}_1 &= \mathrm{FIS2}(U, Q) \\ \dot{\Theta}_2 &= \mathrm{FIS3}(V, P) \\ \dot{\Theta}_{so} &= \mathrm{FIS4}(R) \end{aligned} \tag{5}$$

A lot of tests [9] with simulation model correcting some parameters of that model was performed enabling to determine values of the membership functions.

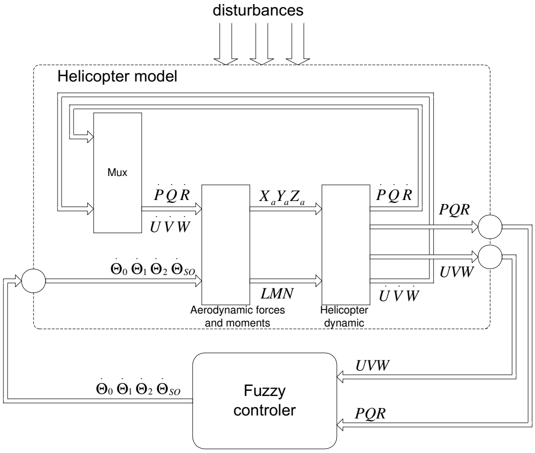


Fig. 1. Helicopter simulation model

6 Results

The simulations were done in two cases: with and without disturbances. Only few results presented in figure 2 (other was not included because of lack of space) confirm efficiency of the proposed techniques of building fuzzy regulator. In figure 2 left plot presents the linear velocity W stabilization and in consequence stabilization of the altitude without disturbances while the right plot presents the same parameter obtained in simulation with disturbances. Without disturbances the periodic moves of the helicopter were suppressed and after $70[s]$ and the amplitudes of corresponding parameters did not outrun values: $U = 1.1*10^{-3}[m/s]$, $V = 0.2*10^{-3}[m/s]$, $W = 5*10^{-3}[m/s]$, $P = 10^{-5}[rad/s]$, $Q = 2*10^{-5}[rad/s]$, $R = 2*10^{-5}[rad/s]$. In practice these values mean that the simulated hover state with application of proposed fuzzy regulator is near the ideal hover state while in the same conditions but without fuzzy regulator the helicopter would go to the crash. In case of restrained blowing presence defined in [10] the stabilization of the periodic small moves of the fuselage was achieved in all cases. However, longer time of parameter stabilization was observed than in case of lack of the wind blowing. Biggest boggles of parameters took place always in the direction of the blow.

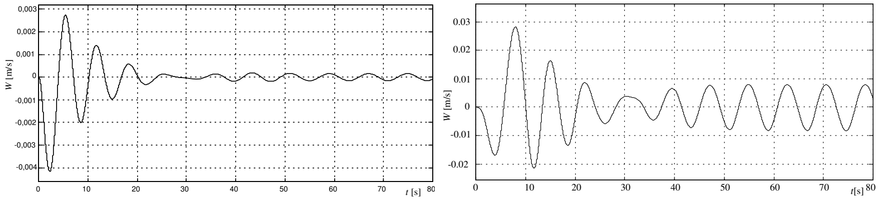


Fig. 2. Stabilization of the W parameter, left – without the disturbance, right – with the restrained disturbance

7 Conclusions

In this paper, we have developed stable control system based on a pilot behavior. The control rules are derived from the observation process similar to an educational introduction into pilot age. Moreover simplicity of the model and strait connection between natural language description and fuzzy rules set let us think that other flight states can be develop this way. The improvement of fuzzy modeling for control is possible based on precise identification the pilot decision process. There is a plan to measure reaction of pilot in real system pilot-helicopter. Also values used in membership functions (see Tables 1 and 2) have to be optimized e.g. like in [8]. Since flying a helicopter is an extremely cost and difficult task, we need cooperation with others.

References

1. P. Bendotti and J.C. Morris. Identification and stabilization of a model helicopter in hover. In *American Control Conference ACC'95*, Seattle, USA, (1995)
2. A. Hassan and B. Charles. Simulation of realistic rotor blade-vortex interactions using a finite-difference technique. *Journal of the American Helicopter Society*, 36(3) (1991)
3. R. A. Hess and C. Gao. A generalized algorithm for inverse simulation applied to helicopter maneuvering flight. *Journal of the American Helicopter Society*, 38(4) (1993)
4. K. Hirota. *International Application of fuzzy Technology*. Springer-Verlag, (1993)
5. B. Kadmiry and D. Driankov. A fuzzy flight controller combining linguistic and model-based fuzzy control. *Elsevier Fuzzy Sets and Systems*, **146** (2004) 313–347
6. G.J. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice-Hall, Englewood Cliffs NJ, USA, (1995)
7. G. Kowaleczko, Z. Dzygadło, and Cz. Szendzielorz. Analysis of a helicopter's spatial motion dynamics and helicopter-auto pilot system. projekt badawczy KBN 9 S604 008 07, Technical University of Warsaw, (1997)
8. R. Khosla *Engineering Intelligent Hybrid Multi-agent systems*. Kluwer Academic Publishers, (1997)
9. Michał Lower. *Fuzzy Logic Control of a Non-Linear Object Related to a Helicopter*. PhD thesis, Wroclaw University of Technology, (2000)
10. D. G. Papanikas, A. J. Spyropoulos, and other. Helicopter rotor downwards calculation using the vortex element method for the wake modelling. *20th Congress of the International Council of the Aeronautical Sciences ICAS*, 1.2.1 (1996)
11. R. Witkowski. *Budowa i pilotaz smigłowców (ang. Instruction to pilotage of helicopter)*. WKiŁ, Warszawa (1979)

Author Index